Metagenomic Analysis of Lung and Oral Microbial Communities Using Whole Genome Shotgun Sequencing

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät
der

Universität Zürich

von

Rounak Feigelman

aus Indien

Promotionskomitee

Prof. Dr. Christian von Mering (Vorsitz)

Prof. Dr. Wolf-Dietrich Hardt

Prof. Dr. Kentaro Shimizu

Zürich, 2016

ॐ भूर्भुवः स्वः

तत्सवितुर्वरेण्यं

भर्गो देवस्य धीमहि

धीयो यो नः प्रचोदयात्॥

Abstract

A vast number of microbes co-inhabit the human body at a wide range of anatomical sites where they engage in a multitude of tasks critical to human health. Among other roles, they provide resistance against invasive microbes, aid in digestion of nutrients and modulate behavior via the production of signaling molecules. Despite their crucial role in maintaining the normal health, microbial communities have only been investigated to a limited extent due to the difficulty of cultivating the majority of microbes outside of their natural environments. However the advent of next generation sequencing, increased computing power and vast storage systems over the past decade have provided the opportunity to study the human microbiome with unprecedented resolution, leading to a steep increase in the number of metagenomic studies centered around healthy and diseased human microbiome. As a result we now know that many complex infectious diseases arise due to compositional shifts in the endogenous microbial community rather than as a result of colonization by a single pathogen. Thus it is necessary to derive a detailed characterization of the complete microbial community rather than single pathogens in order to design effective management and treatment strategies.

In this work, we use high throughput sequencing to draw detailed taxonomic and functional characterizations of microbial communities present in diseased and healthy individuals. To gain this insight, we examine the microbial communities of two distinct anatomical sites. In the first study, we investigate the lung microbial community of cystic fibrosis (CF) patients with chronic respiratory infections, to determine the community composition and identify specific genomic adaptations that microbes undergo during the course of an infection. Crucially, we use total DNA from expectorated sputum without prior depletion of human DNA to provide a comprehensive view of the airway flora. We perform strain typing of abundant lung colonizers to provide a high resolution taxonomic characterization, using well populated multi locus sequence typing (MLST) databases. This method provides a less cumbersome alternative to classical MLST for strain identification and tracking pathogen evolution during disease progression. We show that genomic analysis in conjugation with routine clinical diagnostics can help identify emerging, previously unidentified pathogens and distinguish them from other closely related species. Interestingly, we identify a likely new *Achromobacter* species in a CF lung which can be easily mistaken for its close relative *A. xylosoxidans*, an opportunistic CF pathogen, in the absence of such genomic analysis. We further study strain specific adaptations present in a multidrug resistant (MDR) opportunistic pathogen, *S. maltophilia*, by reconstructing

its near complete genome. Identification of new adaptations not only allows tracking pathogen evolution but can also provide a unique opportunity to identify unique targets for designing anti-virulence drugs for effective management of otherwise resilient MDR microbes, especially in patients suffering from chronic infections. In this specific MDR microbe, we identified the presence of a recently incorporated virulence associated secretion system in its CF-associated strain. We also observed the presence of antibiotic resistance conferring gene sets by analyzing the lung metagenome of CF patients and found good agreement between clinical and genomic predictions. These findings suggest that routine metagenomic surveys of CF sputum microbiome can provide complementary information to routine clinical investigations to prepare extremely detailed infection histories and design more informed and effective treatment strategies.

Outside of infection, microbial communities associated to different body sites evolve dynamically over extended periods of time. To characterize the evolutionary aspect of microbial communities, we turned to the ancient oral microbiome which we studied using high throughput sequencing on two calcified dental plaque samples (known as dental calculus). Interestingly, we were able to identify several commensals and opportunistic pathogens implicated in causing cardiovascular, upper respiratory and oral diseases. While most commensals detected in the ancient plaque are routine members of the contemporary oral community, the presence of opportunistic pathogens in the former was indicative of poor oral health. The assembled genomes allowed us to investigate an ancient strain of a major periodontal pathogen called *T. forsythia*, wherein we identified genomic changes and found evidence supporting lack of pathogenicity related genomic islands present in the modern reference strain. On further functional characterization of the ancient oral metagenome, we detected the presence of several putative antibiotic resistance-conferring gene sets. Surprisingly, this finding indicates that presence of antibiotic resistance mechanisms in microbes predates the use of antibiotics. Lastly, we also identified a few sequences with homology to other eukaryotic genomes of most likely dietary origin. Thus it can be safely concluded that a lot can be learned about the ancient oral health, microbiome and diet by analyzing dental calculus.

These two analyses have highlighted the power applying metagenomic sequencing to non-invasive samples such as sputum or dental plaque, and the wealth of information that can be recovered with relative ease. These methods and the data they can deliver will be very valuable for the development of personalized healthcare strategies and improved routine health management practices for healthy individuals.

Zusammenfassung

Eine grosse Anzahl an Mikroben bewohnen die verschiedensten Körperregionen des Menschen, und tragen mit einer Vielzahl von Funktionen zur Gesundheit des Menschen bei. Unter anderem bieten sie Widerstand gegen invasive Mikroben, helfen bei der Verdauung von Nährstoffen und beeinflussen das Verhalten anderer Mikroben durch die Herstellung von Signalmolekülen. Trotz ihrer entscheidenden Rolle für die Gesundheit des Menschen, wurden mikrobielle Gemeinschaften bisher nur in begrenztem Umfang untersucht, da Kultivierungsmöglichkeiten ausserhalb der natürlichen Habitate für die meisten Mikroben fehlen. Allerdings haben die Entwicklung von neuen Sequenzierungsmethoden (NGS), schnelleren Computern und grösseren Speichersystemen die Untersuchung des menschlichen Mikrobioms in einer bisher unerreichten Auflösung ermöglicht. Diese Entwicklungen haben zu einem Anstieg von publizierten metagenomischen Studien über das Mikrobiom von gesunden und kranken Menschen geführt. Einige dieser Studien konnten zeigen, dass viele komplexe Infektionskrankheiten durch Veränderungen in der Komposition der endogenen mikrobiellen Gemeinschaft entstehen und nicht etwa als Folge einer Kolonisation durch einen einzigen Krankheitserreger. Daher ist es notwendig, eine detaillierte Charakterisierung der vollständigen mikrobiellen Gemeinschaften durchzuführen, um effektive Behandlungsstrategien zu etablieren, anstatt einzelne Krankheitserreger zu studieren.

In dieser Arbeit haben wir High-Troughput Sequenzierung verwendet, um eine detaillierte taxonomische und funktionelle Charakterisierung mikrobieller Gemeinschaften in kranken und gesunden Menschen zu erreichen. Insbesondere wurden mikrobielle Gemeinschaften in zwei anatomischen Körperregionen untersucht. In der ersten Studie inspizierten wir die mikrobielle Gemeinschaft der Lunge von Mukoviszidose-Patienten (auch zystische Fibrose, ZF genannt) mit chronischer Infektion der Atemwege. Wir bestimmten die mikrobielle Zusammensetzung und identifizierten spezifische genomische Anpassungen der Mikroben im Verlauf einer Mukoviszidose-Infektion. Wir analysierten die gesamte DNA in abgehustetem Sputum, inklusive menschlicher DNA, und erhielten eine umfassende und unvoreingenommene Sicht auf die Atemwegsflora. Für eine Stammtypisierung und taxonomische Charakterisierung der mikrobiellen Kolonien wurden etablierte multi-locus sequence typing (MLST) Datenbanken verwendet. Diese Datenbanken ermöglichen eine weniger umständliche Alternative zur klassischen MLST und erlauben die Stammidentifizierung und Entwicklungsverfolgung von Erregern während einer Erkrankung, ohne den Einsatz von stammspezifischer Reagenzien. Wir konnten zeigen, dass genomische Analysen in Verbindung mit klinischen Routinediagnostiken

unbekannte Erreger identifizieren und von anderen eng verwandten Arten differenzieren können. Unsere Studie führte zu der Entdeckung einer neuen *Achromobacter* Spezies in einem Mukoviszidose-Patienten, welche in Abwesenheit von Genomanalysen mit der eng verwandten Spezies *Achromobacter xylosoxidans* hätte verwechselt werden können. Des Weiteren untersuchten wir spezifische Anpassungen im Genom des multiresistenten (MDR), opportunistischen Erregers *Stenotrophomonas maltophilia*. Für diese Arbeit wurde nahezu das gesamte Genom des Erregers rekonstruiert. Die Identifizierung neuer Anpassungen ermöglichte nicht nur das Verfolgen der Evolution des Erregers, sondern bietet auch eine einmalige Gelegenheit spezifische Angriffsziele für antivirulente Medikamente zu bestimmen. Dies ist insbesondere bei Patienten mit chronischen Infektionen von grosser Bedeutung, die oft an widerstandsfähigen MDR Mikroben erkranken. In diesem Fall konnten wir eine in jüngster Vergangenheit ins Genom eingebaute Virulenz ermitteln, welche mit dem Sekretionssystem des nahverwandten ZF-Stammes assoziiert ist. Durch die Analyse des Lungen-Metagenoms von ZF-Patienten haben wir auch die Anwesenheit von Antibiotikaresistenz-verleihenden Gen-Sets beobachten und eine gute Übereinstimmung mit klinischen und genomischen Vorhersagen erzielen können. Diese Ergebnisse legen nahe, dass regelmässige metagenomische Untersuchungen von Mikrobiomen im ZF-Sputum wichtige und ergänzende Informationen zu den klinischen Routineuntersuchungen bieten können und detaillierte Infektionsverläufe bereitstellen, um informative und effektive Behandlungsstrategien anzuwenden.

Neben Infektionen entwickeln sich mikrobielle Gemeinschaften in verschiedenen Körperstellen auch dynamisch über längere Zeiträume. Um den evolutionären Aspekt der mikrobiellen Lebensgemeinschaften zu charakterisieren, haben wir das Mikrobiom der Mundflora von sterblichen Überresten von Menschen aus dem 11. Jahrhundert untersucht. High-Troughput Sequenzierung wurde angewandt, um die DNA aus zwei Zahnsteinproben zu inspizieren. Interessanterweise konnten wir mehrere kommensale und opportunistische Pathogene identifizieren, welche dafür bekannt sind, Herz-Kreislauf- und Atemwegs- und Munderkrankungen zu verursachen. Während die meisten kommensalen Pathogene übliche Mitglieder der zeitgenössischen oralen Gemeinschaft sind, ist die Anwesenheit von opportunistischen Erregern ein Zeichen für schlechte Mundhygenie. Die zusammengesetzten Genome ermöglichten uns, den alten Stamm des parodontalen Erregers *Tannerella forsythia* zu untersuchen. Wir fanden heraus, dass die fehlende Pathogenität auf genomische Veränderungen zurückgeführt werden kann, welche mit genomischen Inseln in modernen Referenzstämmen in Verbindung steht. Bei der weiteren funktionellen Charakterisierung der alten Zahnsteinproben haben wir das Auftreten von mehreren potentiellen Antibiotikaresistenz-verleihenden Gen-sets beobachtet. Überraschenderweise zeigten diese Ver-

suche, dass das Auftretten antibiotischer Resistenzmechanismen in Mikroben älter ist als die medizinische Verwendung von Antibiotika. Zusätzlich konnten wir Genomsequenzen identifizieren, welche eine Homologie zu anderen eukaryotischen Genomen besassen und auf aufgenommene Nahrungsbestandteile zurückgeführt werden können. Zusammenfassend können wir festhalten, dass die Mundgesundheit, das Mikrobiom und die Ernährung von Menschen aus früheren Zeitepochen mit Hilfe von Zahnstein analysiert werden können.

Beide Studien unterstreichen die Leistung metagenomischer Sequenzierungsmethoden für die Gewinnung von wertvollen Informationen durch nichtinvasive Proben wie Sputum oder Zahnbelag, die verhältnismässig leicht gewonnen werden können. Methoden und Daten welche diese und ähnliche Studien liefern, sollten in naher Zukunft zur Entwicklung von personalisierten Gesundheitsstrategien und verbesserten Routinen im Gesundheitsmanagement führen.

Contents

List of Figures 9

List of Tables 11

List of Publications 13

I Introduction 15

1 Microbiomes 17
- 1.1 Microbes . . . 17
- 1.2 The human microbiome . . . 19
  - 1.2.1 Oral microbiome . . . 22
  - 1.2.2 Lung microbiome . . . 24
  - 1.2.3 Gut microbiome . . . 26

2 Classical and modern investigative methods for studying microbial communities 29
- 2.1 Culture-based techniques . . . 29
- 2.2 Culture-independent molecular techniques . . . 30
- 2.3 DNA sequencing . . . 30
  - 2.3.1 Sanger sequencing . . . 31
  - 2.3.2 Next generation sequencing . . . 32
- 2.4 Downstream analysis of metagenomic data . . . 34

3 About this thesis 39
- 3.1 Contributions . . . 39
- 3.2 Challenges . . . 40

II Results 41

4 Sputum DNA sequencing in cystic fibrosis: non-invasive access to pathogen genome information and strain identity 43
- 4.1 Abstract . . . 43

4.2 Background 44
4.3 Results and discussion 46
4.4 Conclusions 58
4.5 Methods 59

5 Pathogens and host immunity in the ancient human oral cavity 69

6 Other contributions 83

7 Concluding Remarks 85

III Back Matter 89

Bibliography 91

Appendix: Sputum DNA sequencing in Cystic Fibrosis: non-invasive access to pathogen genome information and strain identity 107

Acknowledgements 120

List of Figures

1.1 Changes in human gut microbiome with age and extrinsic factors 20
1.2 Relationship between host, microbime and infectious diseases . 21
1.3 Chronic lung infection and CF progression 25

2.1 Workflow of sequencing technologies 32
2.2 Metagenomcis project 35

4.1 Project workflow and sample characterization 46
4.2 Entropy landscape plots 50
4.3 Strain typing for *S. maltophilia* isolate 52
4.4 Strain typing for *Achromobacter* isolate 53
4.5 mucA gene alignment 56
4.6 Comparison of CF-00 *S. maltophilia* isolate against database strains 57

7.1 Sequencing throughput 107
7.2 Non-human DNA vs total DNA 108
7.3 Shannon entropy of subject groups 109
7.4 Taxonomic composition of subject gropus 110
7.5 Entropy plot for CF-99 and CD-34 115
7.6 Entropy plot for CD-42, CD-54, H-84 116
7.7 Entropy plot for H-380, H-94, S-81 117
7.8 Entropy plot for S-82 118
7.9 *S. maltophilia* species tree 119

List of Tables

2.1 NGS specifics 33

4.1 Clinical and genomic antibiotic resistance prediction for CF-85 . 54
4.2 Summary of genomic and clinical predictions 55

7.1 Clinically and genomically detected microbes 107
7.2 CF demographics data 111
7.3 Antibiotics sestivity results for CF-76 112
7.4 Antibiotics sestivity results for CF-00 112
7.5 Antibiotics sestivity results for CF-82 112
7.6 Antibiotics sestivity results for CF-94 113
7.7 Antibiotics sestivity results for CF-99 113
7.8 Antibiotics sestivity results for CF-992 114

List of publications

Manuscript: Rounak Feigelman, Christian R. Kahlert, Florent Baty, Frank Rassouli, Rebekka L. Kleiner, Philip Kohler, Martin H. Brutsche, Christian von Mering. Sputum DNA sequencing in Cystic Fibrosis: non-invasive access to the lung microbiome and to pathogen details. (submitted) *BMC microbiome* (2016)

Paper I: Christina Warinner, Joao F Matias Rodrigues, **Rounak Vyas**, Christian Trachsel5, Natallia Shved, Jonas Grossmann, Anita Radini, Y Hancock, Raul Y Tito, Sarah Fiddyment, Camilla Speller, Jessica Hendy, Sophy Charlton, Hans Ulrich Luder, Domingo C Salazar-Garcia, Elisabeth Eppler, Roger Seiler, Lars H Hansen, Jose Alfredo Samaniego Castruita, Simon Barkow-Oesterreicher, Kai Yik Teoh, Christian D Kelstrup, Jesper V Olsen, Paolo Nanni, Toshihisa Kawai, Eske Willerslev, Christian von Mering, Cecil M Lewis Jr, Matthew J Collins, M Thomas P Gilbert, Frank Ruhli & Enrico Cappellini. Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics* 46, 336344 (2014)

Paper II: Lisa Maier, **Rounak Vyas**, Carmen Dolores Cordova, Helen Lindsay, Thomas Sebastian Benedikt Schmidt, Sandrine Brugiroux, Balamurugan Periaswamy, Rebekka Bauer, Alexander Sturm, Frank Schreiber, Christian von Mering, Mark D. Robinson, Barbel Stecher, Wolf-Dietrich Hardt. Microbiota-Derived Hydrogen Fuels *Salmonella* Typhimurium Invasion of the Gut Ecosystem. *Cell Host & Microbe*, 14(6), 641-651 (2013)

Paper III Tamas Dolowschiak, Anna Angelika Mueller, Lynn Joanna Pisan, **Rounak Feigelman**, Boas Felmy, Mikael Erik Sellin, Sukumar Namineni, Bidong Dinh Nguyen, Sandra Yvonne Wotzka, Mathias Heikenwaelder, Christian von Mering, Christoph Mueller, Wolf-Dietrich Hardt. IFN-γ hinders recovery from mucosal inflammation during antibiotic therapy of *Salmonella* gut infection, *Cell Host & Microbe*, In Press (2016)

Part I

Introduction

Chapter 1

Microbiomes

1.1 Microbes

Microbes are present everywhere in nature. They can be found at the bottom of the oceans [55], in soil [129], hot springs [138], on plants [113] and even in our intestines [77]. Microbes are the most ancient and primitive form of life present on Earth. They have adapted to survive in extreme conditions and can derive nutrition from unlikely sources such as antibiotics [26]. Historically, microbes have been studied mainly in the context of human health and disease following the initial discovery by Robert Koch that microorganisms are the underlying cause of several diseases. Over the past centuries, our view on their role has evolved in tandem with our understanding of their purpose. For example, we now know that microbes play an important role in ensuring human health [6], regulating crucial environmental processes such as biological fixation of elemental nitrogen and carbon, bioremediation of contaminated sites by degrading pollutants, transformation of metals [108] and assisting in digestion and assimilation of nutrients in our gut. Without them, life as we know it can not exist.

Microbes seldom occur in isolation. Rather, they almost always exist as a part of a community consisting of diverse members which interact with each other by secreting signaling molecules [88], exchanging genetic material [80] (via horizontal gene transfer) or by competing for niche and resources [52]. A microbial community along with its surrounding environment can be rightly called a dynamic ecosystem in which both the community and its environment co-regulate each other. For example soil associated microbes play an important role in regulating plant biodiversity and vice versa [41]. Studies show that disturbance in this feedback can often result in the invasion of the environment by exotic plants or microbial species [133] thereby disrupting the ecosystem.

Many microbial communities are found in close association with animals [69], plants [113] and humans [132]. The individual community members interact with their host in a variety of ways including mutualistic, commensal or

pathogenic behavior [22]. The spectrum of interactions exhibited by a microbe can be attributed to its capacity to rapidly adapt to any changes in its surroundings. Under stressful conditions, such as low nutrient availability, competition by co-colonizing microbes, osmotic shock or inflammatory attack by hosts immune system, microbes resort to different strategies such as metabolizing alternate sources of nutrition, secreting toxins, altering membrane permeability or becoming dormant to ensure their survival [24]. Such adaptations modify the nature of interaction between community members with their host. For instance, certain fungi such as *Candida* have been observed to commensally inhabit host mucosal surfaces [2]. When this fungus is exposed to environmental stress such as low nutrient availability, it expresses a certain class of adhesins and hyphal associated genes that facilitates its ability to outcompete the co-inhabiting microflora [2]. During this period, it also causes substantial damage to the host epithelium and transitions from being a commensal to a pathogen.

To understand the complex relationship between microbes and their hosts, we need to study them with respect to their communities instead of in isolation. For example, certain bacteria are only pathogenic in the presence of particular interacting species [48]. However, characterizing a community is an extremely challenging task since this includes identifying all the community members, investigating their biotic and abiotic interaction partners and understanding the nature of these interactions. Traditional methods for identifying microorganisms involve isolation of microbes from their natural environments and cultivation under controlled conditions in a laboratory. While this method has been used to study a wide variety of pathogens, most environmental microbes do not grow in isolation [62] limiting the applicability of this method in exploring environmental microbial communities.

Modern techniques for identifying the bacterial fraction of a microbial community rely on sequencing a relatively conserved gene found across all bacteria such as the 16S rRNA gene [72]. This allows identifying known bacterial genera and also annotation of new, previously unobserved ones. This method is widely used to study the richness and composition of environmental bacterial communities. While it provides a reasonable estimate of the genus level diversity of the bacterial population it does not capture the diversity originating from other members of the community.

Another, relatively recent way of characterizing the composition of a complex sample is by sequencing its entire genomic content [43]. This is known as Whole Genome Shotgun (WGS) metagenomics and has become feasible with the advent of high throughput sequencing technologies. WGS metagenomic sequencing data provides a holistic view of the microbial community as it accounts for all the archeal, viral, bacterial and even the eukaryotic community members and the associated host (if any). Sequencing the entire genomes as opposed to a single marker gene allows for the identification and annotation of microbial

community members and associated gene repertoires. By correlating these genes with their functions, a theoretical metabolic potential of the community can be estimated.

Such metagenomic studies provide deep insight in the functioning of microbial communities and their role in regulating their environment. However, such studies also produce copious amounts of data, thus providing a unique set of challenges. Management of the generated sequence datasets is by no means a trivial task since it requires efficient algorithms for analysis (often requiring supercomputers or clusters) and databases with very large storage systems for storing them. The development of such an infrastructure and software requires dedicated resources with constant maintenance. Additionally, lack of standardization in experimental and downstream analysis renders cross comparison between studies a difficult task. In summary, although WGS metagenomics seems a promising technology to study microbial communities, several challenges need to addressed before it can be routinely applied.

1.2 The human microbiome

Every site of the human body hosts a stunning variety of microbes. These microbes outnumber body cells by a factor of 10 and together comprise the human microbiome [53]. The sheer variety and abundance of the resident microbiota has raised questions about their significance for our existence and has become a deeply investigated area of research. In 2007, the Human Microbiome Project (HMP) was established to study the composition of the residing microbial communities in different body sites of healthy individuals by the United States National Institute of Health [105]. More recently, other international efforts such as Meta-HIT have provided comprehensive characterization of the human gut microbiome in Europeans [42]. These studies have generated large 16S rRNA and WGS datasets providing a preliminary understanding of the function and importance of the human microbiome.

The microbiome is first acquired at the time of birth and continues evolving thereafter [100]. The mode of delivery primarily determines the founding microbiota of the neonates [38], whereafter breastfeeding and diet significantly shape the nascent gut microbiome [45]. As the infants grow older and start interacting with their environment, different body sites acquire characteristic microbial communities that perform a variety of functions crucial to their wellbeing [22], such as aiding in digestion and metabolism of nutrients, resisting colonization by invasive microbes [135] and influencing cognitive behavior [97]. Once fully established, the human microbiome maintains its composition over time in the absence of external perturbations. However, with aging it stabilizes to a new equilibrium. For example, vaginal microbiota have been shown to undergo a shift with menopause [98] and gut microbiota acquire and drop community mem-

bers with increasing age [84]. The human microbiome is routinely disturbed due to daily hygiene practices, change in diet [137] and occasional antibiotic therapies [142]. Figure 1.1 shows changes in the gut microbiome with age, diet, antibitioc use and state of health such as obese, malnourished or healthy. The extent of external perturbations determine the magnitude of change in the community composition. A healthy microbial community is usually not easily perturbed and in the event of a stress episode is able to quickly regain its stable community structure and continue to flourish without any long-lasting effects [22], [21]. Thus resistance and resilience are generally recognized as two characteristic features of a healthy microbial community [5]. However, the resident microbial community can undergo long-lasting changes following repeated antibiotic treatments [31]. Such treatments may also exert selective pressure for resistance phenotypes which have become an increasing cause of concern over the past few decades [8].

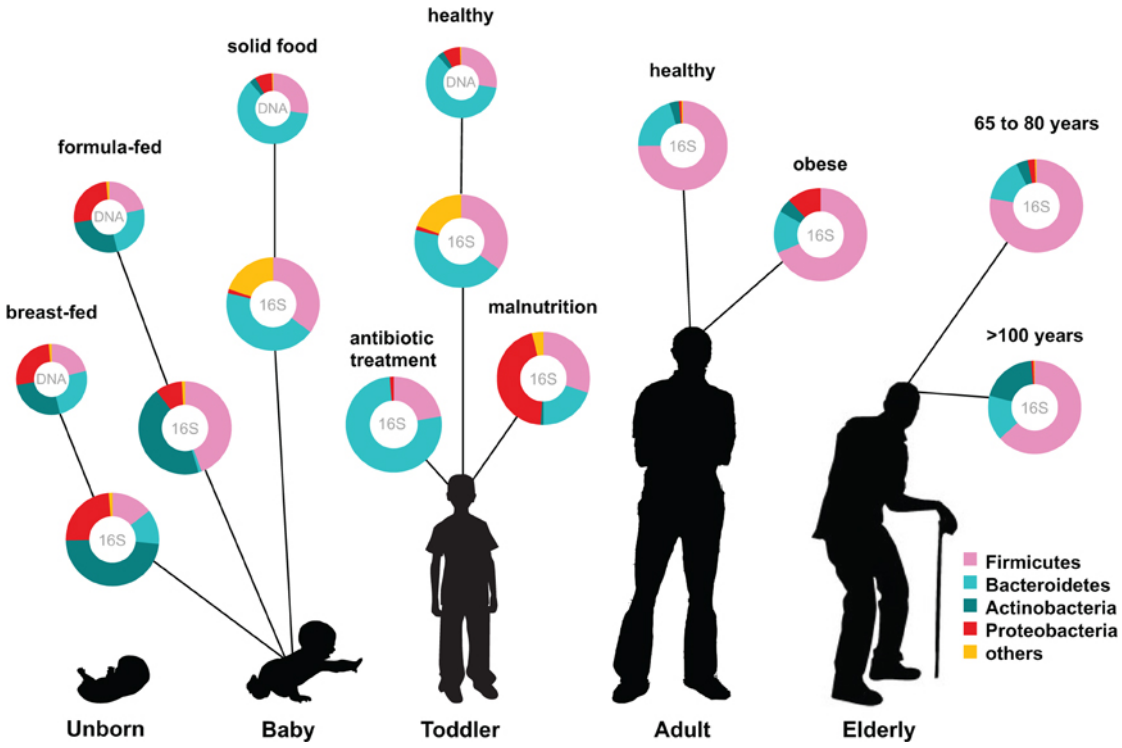

Figure 1.1: Overview of the gut microbiome composition across individuals with differing age, diet, health and antibiotic usage. Diagramatic representation prepared using 16S rRNA and WGS metagenomic studies. Data included for baby from [71] and [117], toddler [89] and [71], adult [14], elderly [14] and [148]. Figure reproduced with permission from [99].

For an external invasive species to colonize a body site, it first needs to breach the immune system, invade the resident microbial community and then successfully compete for primary nutrition sources or metabolize alternate sources [67]. Thus colonization by a pathogen can be viewed as a complex in-

terplay between the immune system, resident microbiota and external invasive species as depicted in Figure 1.2. Opportunistic infections are often facilitated by the frequent use of antibiotics, since they destroy a fraction of the stable microbial community [16] leaving a partially vacant niche that can be easily invaded. For instance, use of certain antibiotics increase the risk of *Clostridium difficile* infections [104]. Alteration in the environment can also result in the establishment of a new species or expansion of a resident community member. For example, Immunocompromised patients or individuals with weakened immunity under stress occasionally suffer from *Candida* and *Aspergillus* infections [81] in the oral cavity resulting from overgrowth of the residing fungal community.

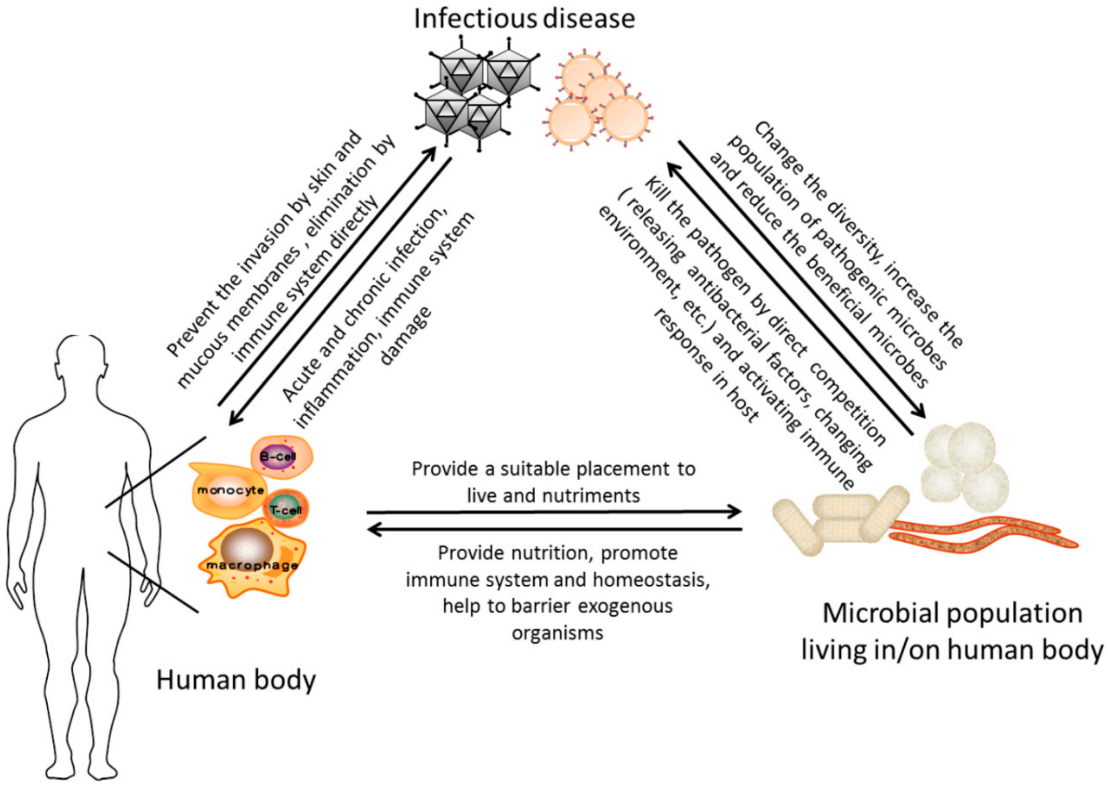

Figure 1.2: Schematic representation of the complex relationship between human host, its microbiome and disease causing pathogens. Figure reproduced with permission from [149]

From an ecological perspective, an infection can be viewed as a state where the endogenous microbial community undergoes a shift due to overgrowth of an existing or a foreign microbe. Although in most cases it is difficult to attribute a causal relationship between changes in the microbial community and a particular disease, functional profiling can nonetheless aid in understanding community dynamics and the role of each associated microbial member during the course of infection. In the following sections the microbiome of three relevant body sites are discussed.

1.2.1 Oral microbiome

The oral microbiome is a rich, spatially heterogeneous microbial community with diversity only second to that of the gut microbiome [21]. It comprises of microbes that inhabit the different surfaces of the oral cavity, namely mucosa, tongue, palate, gums and teeth. Within this community there is a wide variety of commensal microbes that play an important role in maintaining proper health. These commensal microbes compete with exogenous microbes for available surfaces and nutrients thereby providing resistance against infections [63]. However, under altered environmental conditions these commensal microbes can cause diseases and turn pathogenic. For instance, a diet rich in fermentable carbohydrates along with reduced salivary clearance alters the environment of the oral cavity to favor the growth of acidogenic and acidouric species [126]. These bacteria ferment sugars to produce acidic substrates that facilitate demineralization of the teeth in turn causing dental caries. Recent studies indicate that periodontitis, a common oral disease manifests as a result of a dysbiotic microbial community rather than action of a single bacterial species [57].This highlights the importance of studying microbial communities as a whole rather than individuals microbes. The knowledge of the underlying changes that occur in a microbial community during the course of disease initation and progression can ultimately help in developing better routine practices to manage oral health more effectively.

Most oral microbes exist as a part of biofilms that adhere to different sites in the oral cavity [63]. These biofilms are comprised of highly organized collections of distinct microbial species that are embedded in an extracellular polysaccharide matrix [112]. Their composition and structure is primarily determined by local heterogeneity of the oral cavity such as difference in salivary clearance, pH, oxygen content and presence of immunological molecules [3]. Saliva carries nutrients, immunoglobulins and additionally transports bacteria across different oral surfaces, which may lead to the formation of microheterogeneity [3], [49]. Extrinsic factors such as diet, dental hygiene and smoking are some of the factors that shape the biofilm structure and composition [49]. High sugar content food and tobacco smoking have both been shown to alter the composition of oral biofilms in favor of pathogens [63], [86]. Within biofilms, microbes engage in a number of synergistic activities such as quorum sensing and cross feeding [112] which facilitate their growth and survival. Bacteria use quorum sensing to communicate and coordinate their behaviour and benefit from collective action. Microbes may also engage in quorum sensing to gain competitive advantage over other co-inhabiting species. For instance, the pathogen responsible for causing dental caries uses quorum sensing to secrete bacteriocins which gives them a competitive edge within their niche [59], [146]. Due to the close proximity of microbes in a biofilm they frequently engage in the

exchange of genetic material via horizontal gene transfer [112]. This plays an important role in the spreading of antibiotic resistance conferring genes within the oral microbial community. Oral biofilms have been demonstrated to harbor a wide range of resistance genes even in individuals with no previous exposure to antibiotics [20]. This could possibly be due to regular exposure to environmental microbes which results in the introduction of new genes into the microbial community [103]. These genes are then propagated via horizontal gene transfer. For instance, metagenomic studies aimed at assessing the spread of antibiotic resistance genes in the oral microbiome have identified a novel antibiotic resistance mechanism against tetracycline in uncultivable bacteria, indicating a role of the oral microbiome as a reservoir of resistance genes [34].

Occasionally oral microbes get transmitted to different body sites where they establish opportunistic infections. For instance, bacterial pneumonia is very often caused by the aspiration of oral periodontal pathogens into the lower respiratory tract where they establish an infection as a result of reduced mucociliary clearance as is the case of cystic fibrosis patients [4]. Periodontitis has also been associated to an increased incidence of atherosclerosis [87] (excessive thickening of arterial walls). Although the underlying mechanisms for this are not clearly understood, it's been observed that individuals with periodontitis are 400 times more likely to suffer a stroke than with individuals with caries or other oral diseases [56]. In extreme cases, oral microbes have been documented to enter the bloodstream through mucosal lesions and cause severe infections at distant organs such as the liver and the brain [25], [116]. Since oral infections are regularly observed to cause a number of systemic diseases especially in immunocompromised patients or in patients with preexisting conditions, it is important to characterize its members and their role in establishing these infections to develop more informed treatments.

Recent investigations using 16S rRNA studies have shown that there are approximately 2000 bacterial phylotypes present in the healthy oral cavity [32]. Online databases such as the Human Oral Microbiome Database (HOMD) [32] aim to provide a taxonomic schema for the unnamed bacterial phylotypes identified in the oral community. Approximately half of the oral bacterial members can not be currently cultivated in the laboratory [32]. Thus direct sequencing approaches can serve as a powerful tool to characterize the uncultivable fraction of the oral community.

A recent single cell genomics study performed full length genome comparison of an uncultivable oral commensal *Tanerella* BU063 with its closest known relative *Tanerella forsythia*, a member of the red complex (group of periodontal pathogens) [11]. This study detected important differences in metabolic pathways, loss in synteny and absence of several virulence associated genes in BU063 supporting its commensal nature [11]. Such genomic studies provide deep insights into the evolution of pathogenicity and difference in the nature of

host microbial interactions among closely related pathogenic and commensal strains in the oral cavity [11]. Thus direct sequencing approaches can provide crucial information about oral microbes including the unobserved community members.

1.2.2 Lung microbiome

Historically, healthy lungs were thought to be sterile [10]. This notion has been revised in the past decade by several high throughput sequencing studies that detected the presence of a robust microbial community in the healthy respiratory tract [44]. For a long time, recovery of oral bacterial community members from lung samples was presumed to be contamination during sampling but recent studies have shown that the healthy lungs harbor a microbial community founded in part by the oral microbiota [19]. While the community composition has been shown to differ, the strong overlap in the community members is due to the continuity of the airways. In addition to the oral microbes that are aspirated into the lungs, the immune system, gut microbiome [96] and the surrounding environment [136] are some of the other factors that affect the composition of the lung microbial community. Several studies have tried to account for the effect of these interaction variables [40] but it has been a challenging task to decipher their relationships.

Beginning from the nasal passage to the alveoli, the healthy respiratory tract offers a warm and moist environment for microbes to colonize [35]. However, he presence of cilia on the respiratory epithelium deters microbes from adhering [15] and the bacteriostatic mucus lining [143] prevents them from growing. These mechanisms assist in regulating the microbial load in the airways. Lungs also exhibit substantial spatial heterogeneity in its regional growth conditions such as temperature, oxygen availability, pH and host epithelial interactions [139]. This spatial heterogeneity translates only modestly into the formation of localized communities [35] emphasizing that other factors such as bi-directional sweeping of the microbes in airway plays a greater role than localized selection pressures in shaping the community composition [36]. Any disturbance in these regulatory factors is bound to alter the resident microbial community composition.

The majority of respiratory diseases have an infectious component, including chronic obstructive pulmonary disease (COPD), cystic fibrosis (CF) and chronic bronchitis [12]. These diseases manifest as an interplay between environment, immune system, genetic predispositions and an imbalance in the residing microbial community. The lung microbiota of patients suffering from CF or COPD shows a markedly different composition as compared to healthy individuals. Altered lung environment such as reduced mucociliary clearance in most cases [37] facilitates the onset of bacterial and fungal infections. Clogging of the

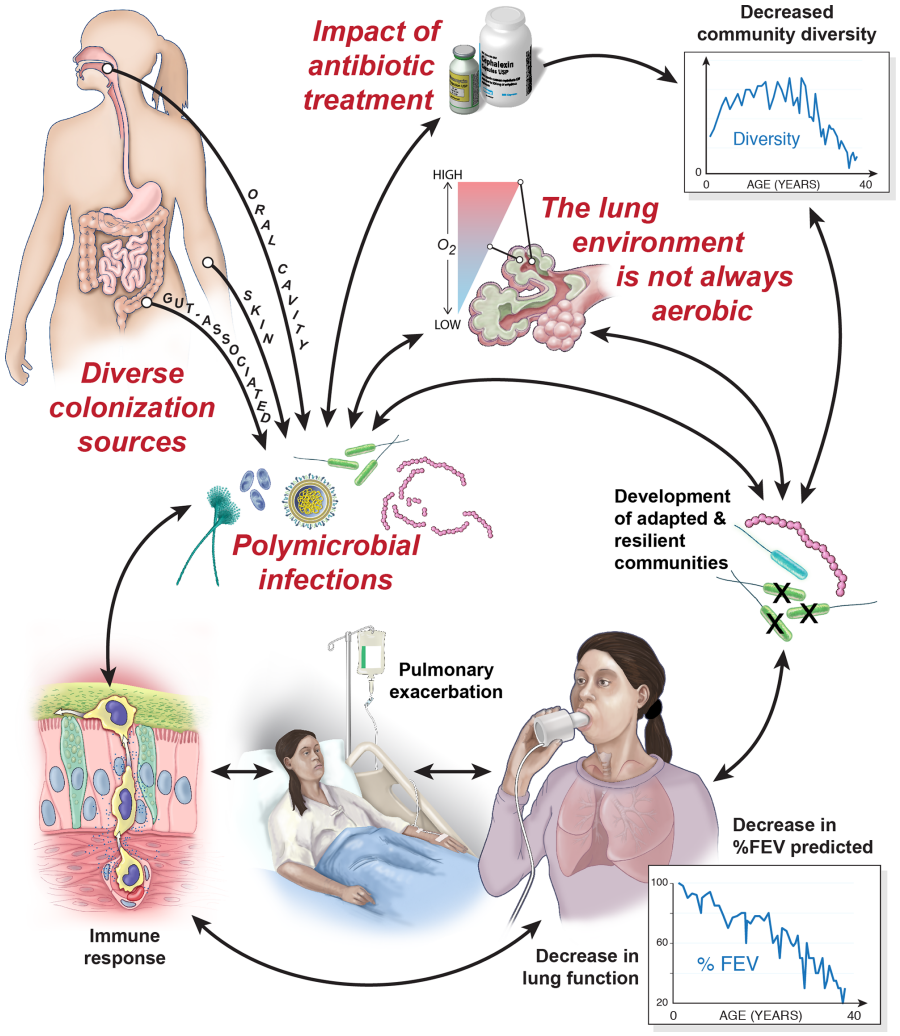

Figure 1.3: Pictorial representation of the various factors involved in recurring, chronic lung infections. Figure reproduced with permission from [73]

airways with mucus secretions creates pronounced spatial heterogeneity (hypoxic environments) which facilitates localized overgrowth of microbes and also makes it harder for antibiotics (administered through inhalation) to take effect [73]. Progressive lung infections have been shown to reduce lung microbial diversity over time [150]. This reduction in diversity has also been linked to lung function decline and poor disease outcome [150]. Figure 1.3 shows that microbes associated to other body sites such as the oral cavity, skin and gut can establish polymicrobial respiratory infections within an altered lung environment. These infections usually trigger an immune response, which besides killing the microbes also causes inflammation (exacerbations) leading to lung function decline. Additionally, repeated antibiotic treatments to manage the microbial load results in loss of lung commensal flora and concomittant development of antibiotic resistant microbial communities. The development of such communities hastens lung function decline and leads to poor disease outcome.

Microbiome investigations are relatively recent in the context of lungs. A majority of respiratory microbiome studies have been centered around investigating diseased states in order to elucidate the role of the lung microbiome in respiratory infections and recovery. However, there is limited knowledge of the core microbial community members and their functions in healthy individuals. Thus, before any microbiome based treatments for managing respiratory infections can be designed, large scale population-based studies are required to characterize the healthy lung microbiome and its dependence on factors such as age, lifestyle and environment.

1.2.3 Gut microbiome

The gastrointestinal tract, more commonly known as the gut, is home to the most diverse and complex microbial community within the human body [32]. Studies suggest that the gut microbiome asserts local as well as systemic effects in maintaining proper health. It does so by playing an important role in the development of adaptive immunity [60], metabolism of otherwise complex nutrient sources [107], biosynthesis of vitamins [74] and neurotransmitters [144], and providing resistance against colonization by pathogens [135]. The current estimates suggest that upto 10^{12} microbial cells colonize per gram of the colon [140].

The gut microbiome composition is influenced by several intrinsic and extrinsic factors including (but not limited to) diet [58], age [134], sex, basal metabolic rate [39] host genotype and medications [61]. For instance, it has been shown that a diet with high fat content alters the gut microbial composition both in the presence and absence of obesity [58]. Besides antibiotics, medication such as corticosteroids profoundly effect the composition of the gut microbiome [61] possibly by affecting the immune system. All these factors directly or indirectly shape the gut microbiome but the extent of their effect has been challenging to assess. Studies carried out on a geographically limited small sample size have in the past suffered from the confounding effects of these interaction variables. However in recent times, studies by the human microbiome project and MetaHit consortium have tried to account for these factors by collecting population wide data. The aim of these large scale studies has been to mainly define a core microbiome associated to the healthy gut (and other body sites), prepare reference genomes of routinely observed members and delineate their taxonomies as well as functional characteristics [106] [42]. The objective is to prepare high quality large datasets and to characterize them for developing a deep understanding of the role played by the gut microbiome in maintaing health [106].

The gut harbors a characteristic microbial community that varies greatly in composition across individuals [132]. These differences have been shown to be slightly reduced between related individuals but even then a consistent, com-

mon set of shared bacterial phyla has not be detected [131]. A group of genes engaged in a core set of metabolic pathways has been observed consistently in the healthy gut microbiome from different individuals [131]. This indicates that while the healthy gut demonstrates large variability at the phyla level, it is functionally conserved at the gene level and engages in similar metabolic activities. Furthermore, deviations from this core functional profile has been observed in individuals with unhealthy or diseased conditions. For instance, obese individuals tend to show decreased microbial diversity and an altered set of core metabolic pathways as compared to healthy individuals [131]. Patients with crohns disease and ulcerative colitis exhibit a greater shift in their functional profile as compared to their microbial composition [90]. This indicates that perhaps a core set of genes and pathways could be a potential alternative to describing a healthy gut microbiome rather than a shared group of microbial species.

Colonization Resistance

A healthy gut microbiome in the absence of any external perturbations such as therapeutics has been shown to inhibit colonization and prevent overgrowth of several enteropathogens. This characteristic of the gut microbiota is known as colonization resistance [135] and has been a topic of investigation of many different studies. It has been shown to do so using a variety of mechanisms which include i. physically colonizing the gut epithelium to reduce anchoring sites for the invasive species [123], ii. competing for nutrient sources iii. stimulating the immune system against the invading microbe [118]. The gut microbial composition has been shown to effect susceptibility of gut to pathogen invasion, in particular communities with reduced species richness being more vulnerable to colonization [122]. While the underlying reasons for this phenotype are not fully understood, it has been shown that certain members of the healthy human gut community can actively repress expression of virulence factors by secretion of metabolites thereby evading infection [29]. Occasionally, despite the resistance mechanisms employed by the gut community, the enteric pathogens become successful in establishing an infection. They do so by circumventing the competition for different carbon sources posed by the endogenous community, by feeding on secondary (i.e. less preferred) carbon sources [65] or showing preferential usage of other nutrient sources as compared to the gut community. Understanding mechanisms of gut colonization by enteric pathogens and the role of gut microbiome in controlling nutrient availability provides a window of opportunity to design methods to manage or even prevent these infections.

Chapter 2

Classical and modern investigative methods for studying microbial communities

Our long standing interest in investigating members of a microbial community has led to development of new technologies that have facilitated identification and characterization of microbial communities and their members. In this section we will discuss some traditional and recent techniques developed for this purpose.

2.1 Culture-based techniques

Standard microbial culture requires isolation and growth of a microorganism on a selective nutrient rich media (most based on an agarose gel), under controlled clinical conditions to identify and characterize the microbe [70]. To date, this method has been routinely used as a diagnostic tool in clinical microbiology laboratories to identify pathogens. However, when investigating complex microbial communities from environmental samples such as soil, a majority of these microbes can not be cultivated in the laboratory using standard culture techniques [62]. Thus, over time the traditional cultivation techniques have been revised to mimic natural environments using a variety of methods. This includes allowing co-culture of microbes, controlling pH, temperature and oxygen content to imitate their niche variables and finally installing a portion of the actual environment as a nutrient source to facilitate microbial growth using various micro-cultivation techniques [124]. With recent advancements in culture techniques, a higher fraction of environmental microbes can be isolated and grown in the laboratories but even then majority of microbes (around 99%) remain uncultivable [62]. Thus molecular techniques that do not depend on laboratory culture of microbes are more likely to provide a well rounded view of environmental microbial com-

munities.

2.2 Culture-independent molecular techniques

Molecular methods that do not rely on isolation and cultivation of microbes for their identification primarily depend on detecting biomarkers that can be used as a proxy for the presence of the microbe itself. When phylogenetic marker genes such as 16S rRNA is used as a biomarker then the bacterial diversity of the community can be estimated. This is especially useful for assessing the diversity in samples from which these bacterial members can not be isolated and cultivated in laboratory cultures. Total DNA from such samples is extracted and then a particular gene of interest such as 16S rRNA gene (i.e. the biomarker) is then amplified and analyzed using a variety of DNA assays such as Denaturing- gradient gel electrophoresis [92] (DGGE), Temperature gradient gel electrophoresis (TGGE), Random amplified polymorphic DNA (RAPD) [51] and Terminal restriction fragment length polymorphism (T-RFLP) [127] to extrapolate diversity of the whole community from their readouts. This process is commonly known as community fingerprinting and is used routinely to track temporal changes in the bacterial diversity of an environmental site.

An important limitation of all these methods is that they cannot be used to identify the bacterial members of the microbial community. Moreover, these assays can be difficult to cross compare when generated from different laboratories and also due to difference in their sensitivity towards lowly abundant microbes [13]. Since these methods are relatively straightforward and inexpensive they serve as a reasonable option for rapidly assessing the community diversity before planning elaborate experiments. In recent times, as the DNA sequencing is getting progressively cheaper, amplicon based sequencing is being more widely adopted for performing diversity analysis and identifying the bacterial community members.

2.3 DNA sequencing

Sequencing is a method used to detect the order of nucleotides on a DNA strand. To perform sequencing, first the DNA needs to be extracted from different samples which can be pure isolates of microbes, environmental samples such as soil from a farm field, water from the tap, or clinical specimens as sputum or stool. Once the DNA is extracted, sequencing is performed either on the entire DNA content or only on a specific genomic region depending on the type of research question being addressed. For instance, when the aim is to investigate the diversity of the bacterial fraction of the microbial community, delineate their taxonomies or deduce phylogenetic relationships then a

universally present gene such as the 16S rRNA found in all bacterial genomes is sequenced [50]. This gene serves as an ideal phylogenetic marker as it is fairly long (1500 base pair) and contains highly conserved regions that are interspersed by varying stretches of DNA. The conserved regions are ideal for designing complementary PCR primers to recover amplicons spanning the variable regions. These amplicons can then be sequenced to recover the readouts. By converting the differences (mismatches) between the variable regions of the 16S rRNA gene into a meaningful distance measure, operational taxonomic units (OTUs) can be drawn which are essentially clusters of highly similar sequences that satisfy a distance cutoff [114]. The number of OTU clusters found in a microbial community can then be used to measure diversity or estimate the number of bacterial genera present in a sample. Alternatively, when the functional potential or whole community structure including fungal, viral and archeal members of the microbial community need to be investigated, the entire genomic content as opposed to a single gene is sequenced. This is known as whole genome shotgun sequencing (shotgun implies sequencing is done on fragmented DNA molecules) and has been used to investigate complete metagenomes to address a wide variety of research questions linked to function and composition of environment or human associated microbial communites.

2.3.1 Sanger sequencing

Sanger sequencing or also known as chain termination method was first developed in 1977 [115]. Since then, it has been widely used for sequencing high quality bacterial and eukaryotic reference genomes. Sanger sequencing was the primary sequencing technique used for preparing the first draft of the human genome. Its long read length $\sim$1000 base pairs and low error rate [119] makes it a gold standard in DNA sequencing technologies. While the accuracy of sanger sequencing makes it an ideal technique for clinical applications, its prohibitively high per base cost of $\sim$ \$0.50/base [119] and low throughput makes it less attractive for conducting exploratory analysis on host and environment associated microbial communities.

In recent times, the original sequencing technique has been sped up by optimizations such as parallelization using 96 well plates and minimizing manual intervention. The original technique has been modified to use dyes for detecting the four di-deoxynucleotides but the sequencing principle is essentially still the same [125]. Figure 2.1 shows the workflow of the modern capillary based high throughput sanger sequencer.

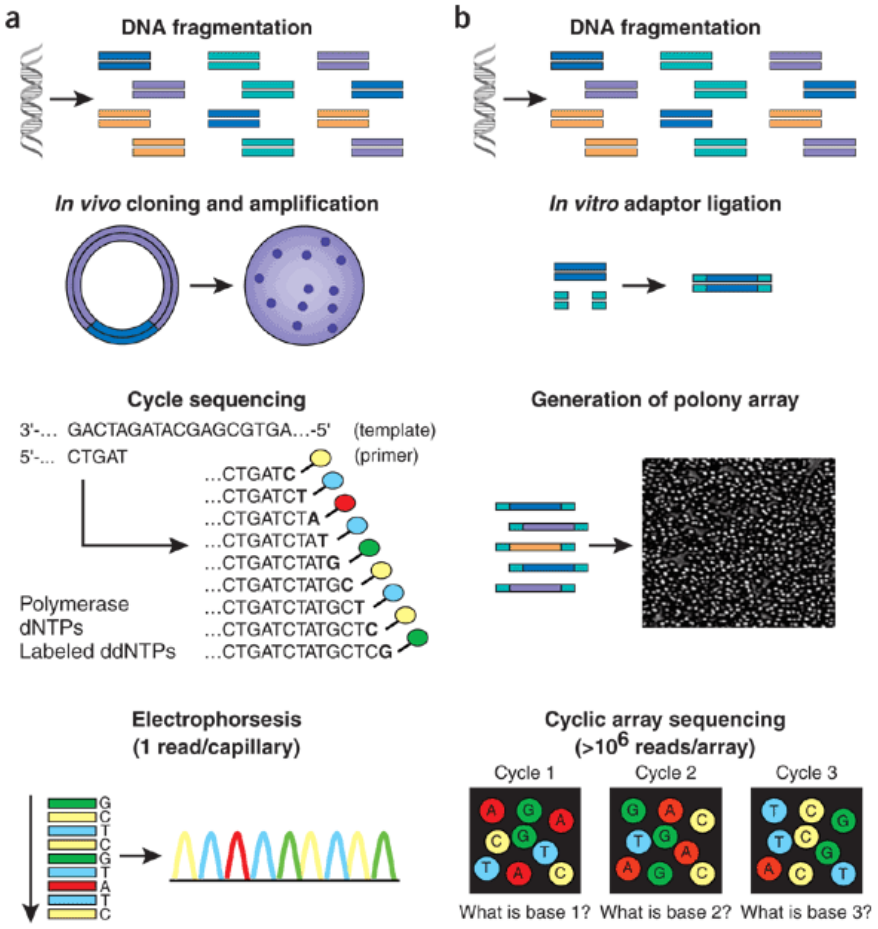

Figure 2.1: Workflow of sequencing technologies. a. Steps involved in capillary based high throughput sanger sequencing. b. General workflow of cyclic array next generation DNA sequencing paltforms. Figure reproduced with permission from [119]

2.3.2 Next generation sequencing

In late 90s, sequencing platforms were designed such that millions of small DNA sequences could be read in parallel. These massively parallel, high throughput sequencing technologies are commonly referred to as next generation sequencing (NGS). One of the first next generation platforms to be commercially was designed by 454 systems [83]. Since then, many sequencers have been commercially released by companies like Illumina [46],[130], Pacific Biosciences [76] and SOLiD [120]. The main features of next generation sequencing that sets it apart from previous methodologies is its low per base sequencing cost and high throughput [119]. Due to these reasons, it has revolutionized research particularly in the field of human health and environmental microbial ecology. With the dramatic reduction in cost, sequencing has become affordable and accessible to a majority of researchers. This has led to a burst in sequencing based projects. Furthermore, different sequencing techniques (based on detection of

different fluorochromes, or in change in pH) have opened avenues to address a broad variety of research questions. Like sanger sequencing, next generation sequencing is also based on the principle of sequencing by synthesis. While there are major differences in the specifics of different next generation sequencing technologies, the basic workflow is largely similar and has been depicted in Figure 2.1. The library preparation for sequencing begins with fragmenting the genomic DNA in smaller pieces in the order of 200-600 base pairs. Following fragmentation, adapters are ligated to both the ends of all the sequences. These adapter ligated sequences are spatially fixed or isolated from one another (depending on the sequencer) and are clonally amplified. Thus all the clones derived from a single DNA sequence are spatially localized on an array. This array of DNA templates is then flooded with a mixture of appropriate enzymes and nucleotides to allow addition of the complementary nucleotide base. During this process a signature signal of the added nucleotide is released. This signal is recorded (often by a very sensitive laser) after each flooding step. The process is repeated to generate readout of desired length. Due to increasing noise to signal ratio, the base quality decreases over time and this determines the length of the read outs in different technologies. Specific platform related details have been summarized in Table 2.1.

Sequencer	Feature generation	Synthesis mechanism	Read (bp)	Error type
454	Emulsion PCR	Pyrosequencing, PCR	700	Inert-deletion
Illumina	Bridge PCR	Reversible terminators polymerase	150*2	Substitution
SOLiD	Emulsion PCR	ligase	60*2	Substitution
PacBio	Single molecule	Polymerase	33	Deletion

Table 2.1: Next generation sequencing technology related specifics. Data adapted from [119]

While next generation sequencing is a promising technology, it has been unable to replace sanger sequencing and establish itself as a reliable technique to draw clinical inferences. This can be attributed to its relatively higher error rate in base calling and shorter read length that necessitates elaborate downstream data processing. A single Illumina Hiseq4000 flowcell generates up to 1.5 terabyte of sequencing data [75]. Processing such large datasets require investment in powerful machines and storage systems and need experts for its analysis. Such elaborate in house setups for data analysis can be prohibitive in the use of next generation sequencing as a routine clinical tool.

With growing interest in this technology, some of the technical shortcomings are bound to be addressed with time. As for the downstream analysis, many

centralized sequencing centers often provide service in data analysis. Though, in some cases this might not be a feasible option with data privacy concerns. Irrespective of its current shortcomings, next generation sequencing has in true sense transformed the study of complex microbial communities since it has enabled their investigation in context of their natural habitats. While high throughput amplicon sequencing studies have shed light on the large bacterial diversity present everywhere, high throughput whole genome shotgun sequencing has facilitated their functional characterization thereby aiding in understanding the role of different community members in the biochemical activities that take place in an environment.

2.4 Downstream analysis of metagenomic data

As previously discussed, large amounts of data can be produced by indiscriminately sequencing the total DNA extracted from a complex sample. The study of this data to gather meaningful information about the residing microbial community and its associated host (if any) is known as whole genome metagenomics [110]. In recent times, there has been a surge in the number of available tools for the analysis of metagenomic sequence data. However, each tool has its pros and cons and underlying assumptions that makes it ideal for some datasets and unsuitable for others. It is therefore important to consider these issues before embarking on data analysis to save time and money since the cost of re-running computationally demanding tools can sometimes be higher the sequencing cost itself [141].

The first step of any biological experiment is sample collection and preparation which is also a crucial step in metagenomic projects. The choice of sample determines the quality of data that can be recovered from it and should be such that it captures the entire community that is to be analyzed. During sample collection, its is highly recommended to record the associated meta information such as sample source and conditions to facilitate data replication. Guidelines such as MIMS (Minimum Information about Metagenome Sequence) [145] aid in comprehensively logging metadata for future reference. Next, depending on the type and form of the sample such as tissue [102], soil [30], dental plaque [101] and stool [121] there are tailored protocols available for robust extraction of DNA. Bead beating has been shown to increase the yield of microbial DNA since it helps in efficient lysis of microbial cell walls [28]. Often host associated samples contain a large amount of eukaryotic DNA which can interfere with the efficient recovery of the microbial DNA. For instance, more than 90% of the recovered DNA from sputum samples is generally of host origin which can interfere with the downstream analysis of the associated microbial fraction. In such cases, DNA depletion protocols that allow removal of host DNA by either selectively targeting eukaryotic DNA [47] or with physical separation of cells us-

ing flow cytometry are generally recommended. If a particular section of the microbial community is of interest then enrichment methods to capture particular cell types should be considered[79]. However, protocols aimed at selectively enriching microbial DNA are also likely to introduce some bias in the data which trickles through the entire downstream analysis, thus they should be used with caution.

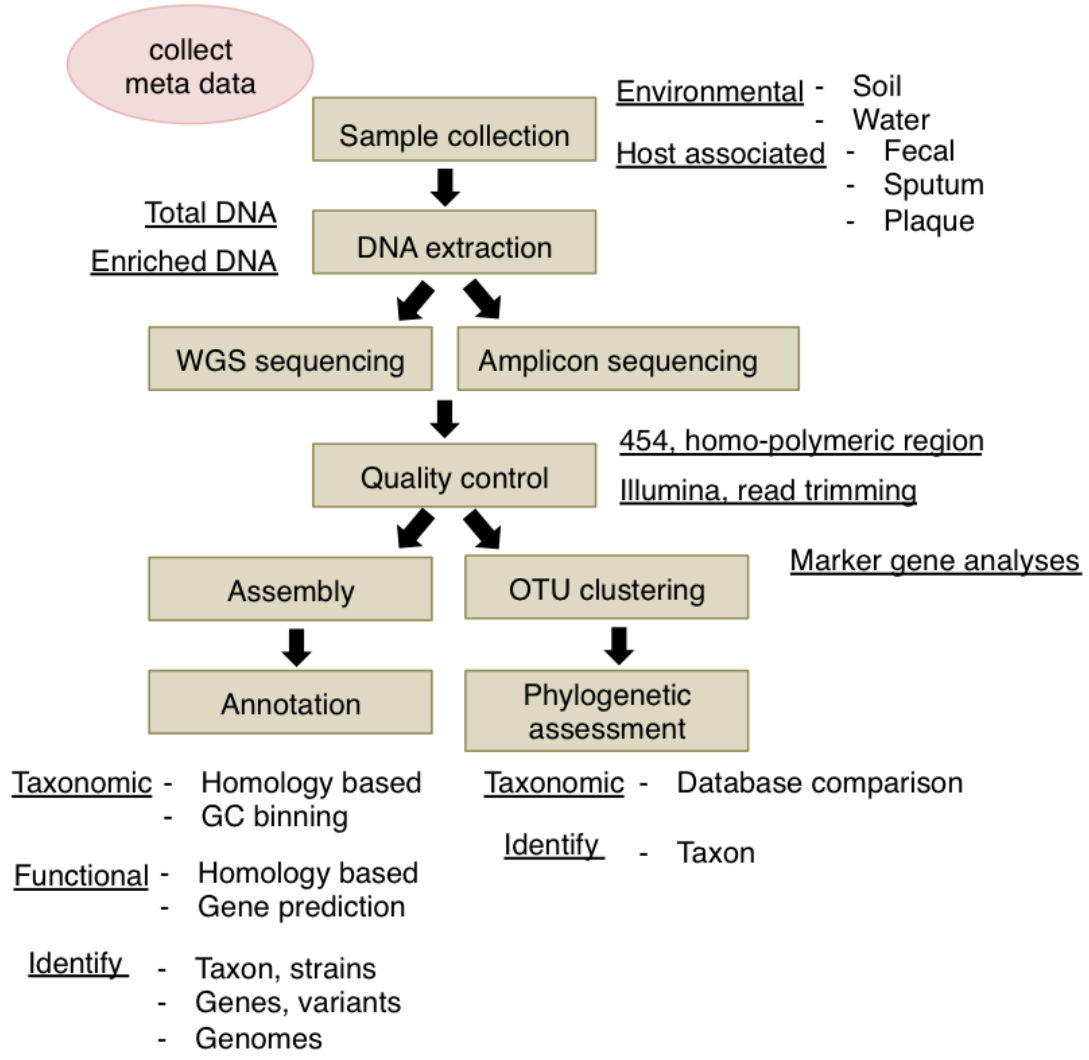

Figure 2.2: Typical workflow of a metagenomics project

Once the DNA is extracted, the next step is to select an appropriate sequencer that returns readouts of appropriate length and acceptable error rate. Since a majority of metagenomic projects generally use either 454 pyrosequencing or Illumina (including projects discussed in this work), these two technologies will be the focus of the following discussion. The choice of the sequencer primarily depends on the research question at hand and this can vary from analysis of taxonomic diversity using 16S rRNA gene region or any other single marker gene, recovery of a near complete genome of an uncultivable mi-

crobial member, identifying polymorphisms in a genomic region, assessing the gene pool of the entire community. 454 sequencer returns reads of up to 800 base pairs which is often sufficient to capture an entire region of a marker gene on a single sequence. This makes it ideal for sequencing amplicon libraries. 454 requires relatively little starting material in the order of a few nanograms for generating single end libraries and has a relatively short run time of a single day [1]. All these characteristics make it an extremely desirable sequencing technology. However, due its base calling method, 454 suffers from an increased error in the form of indels in homopolymeric regions that can lead to frameshifts in coding regions. Since these are non-random errors, they are well characterized and can be identified during data analysis [109]. Additionally, several tools intrinsically account for these errors when dealing with 454 generated sequence data.

Illumina sequencers generate short reads of 150 base pair length with readout of upto 300 base pairs in a paired end mode. Unlike 454, it does not suffer from any systematic errors although the base quality generally drops towards the rear end of the reads [93]. Thus Illumina generated reads might require trimming during quality control step. However, it compensates for the short read length by generating data with high coverage. Illumina sequencers have a throughtput of around $\sim$200 million reads per lane, with usually 8 lanes in a single flowcell. The high coverage also allows multiplexing of samples in a single lane when shallow depth of sequencing is suffecient, catering to multiple samples in a cost effective way. It is by far the cheapest in terms of per base sequencing cost but often requires assembly of short reads into longer contiguous sequences before any reliable conclusions can be drawn from the data. Like 454, it also requires relatively little starting material for library preparation but has a much longer run time of around 11 days [128]. The high coverage sequencing data generated by Illumina sequencers is ideal for calling single nucleotide variations or assessing the entire gene pool of a community. It also allows distinguishing closely related strain types that are often found in metagenomic datasets.

High throughput sequencing data requires assembly i.e. combining overlapping reads originating from a single genomic region into longer contiguous sequences when the aim is to analyze near complete genome of a microbe or recover sequences spanning entire gene sets. There are essentially two ways to approach this issue, namely reference-based or de novo assembly [128]. In reference based assembly, whole genome sequence of a species is used as a reference to assemble genomes of closely related strains present in a metagenomic sample. However, this method is not able to capture synteny (genome rearrangements) and insertions in the new assembled strain. More commonly, reads are assembled using de novo assemblers that construct longer contiguous sequences by identifying overlapping regions in reads without the help of a

reference. These assemblers visualize genomes as directed graphs and construct paths of the graphs by intersecting terminal regions of the reads. Many genomic de novo assemblers, are based on the underlying assumption that sequences belonging to a single genome exhibit uniform coverage and that the microbial species are mostly clonal rather than a cloud of closely related strains that exists together. These assumptions are very often violated in metagenomic datasets and thus many de novo genome assemblers do not perform optimally on metagenomic datasets. However, these assemblers might perform acceptably when there are a few overrepresented, dominating species present in a sample. Examples of some de novo assemblers are SOAPdenovo2 [82] and Velvet [147].

Some genomic assemblers have been further extended to address the above mentioned issues and deal with metagenomic datasets. These include MetaVelvet [94] and MEGAHIT [78]. Although these assemblers are a considerable improvement over the previous generation of de novo assemblers in terms of performnace on metagenomic datasets, more sophisticated algorithms are needed that can assemble reads from lowly abundant microbial species. Empirical tests show that assemblers designed for single cell genomic data such as SPAdes [7] perform well on metagenomic data sets since they perform compound assemblies with varying overlapping lengths (i.e. different k-mer sizes), to deal with non-uniform read coverage.

Once assembled, contigs can be aligned against reference genomes using tools like Mauve [111], [27] or Whole genome Vista [17] that are optimized for whole genome alignments to isolate draft assemblies of closely related strains. Such assemblies are more likely to capture new insertions and genome rearrangements if present in the newly assembled strain. Binning is another approach that can be used to sort sequences into phylogenetically meaningful clusters based on an intrinsic sequence quality such as GC content, representing a taxonomic level such as genus or species. The rationale is that microbial species exhibit a characteristic GC content [91] which is often also represented in long assembled contigs. This method performs well only for long sequences since short sequences are not able to accurately capture this information. Most binning techniques are unable to distinguish strain level diversity, clubbing closely related strains into a single bin that represents a heterogeneous cloud of strains belonging to a species.

After the assembly step, a fraction of the data remains unassembled. Reasons for this might include inadequate sequencing depth or underlying heterogeneity. When the data is dominated by short contigs and unassembled reads then homology based searches are used to divide data into clusters of previously known species.

To gather information about the different features present in the assembled contigs and reads, generally homology based search such as BLAST [18] is

used against a high quality curated dataset. This is often computationally cumbersome and time taking for large datasets. Alternatively, feature prediction tools such as MetaGeneMarK [151] or Metagene [95] can be used to detect coding regions present in an assembled contig. These tools are essentially classifiers that have been trained on codon usage information and initiation sequences to identify protein coding genes. Once the genes have been identified, databases such as EggNOG [64], KEGG [68] and PFAM [9] can be used to assign functional annotation to the identified gene sequences. Besides detecting coding regions, there are several dedicated databases which house high quality ribosomal RNA sequences such as greengenes and ribosomal database project which can be used to detect RNA coding genes sequences present in the dataset.

There have been a number of recent efforts to streamline the process of metagenomic analysis to allow cross comparison between studies. This has resulted into the development of several analysis pipelines that essentially integrate a number of previously discussed tools under a single wrapper. QIIME [66] and MG-RAST [54] and IMG/M [85] are such analysis pipelines that performs OTU construction and diversity analysis and functional assignment on unassembled read data. Additionally, some of these tools provide data repositories to store metagenomic datasets which can be publically shared with the scientific community.

Chapter 3

About this thesis

3.1 Contributions

In this work, I present detailed characterizations of microbial communities associated to two body sites. The sites were investigated in separate studies with the common goal of gaining insight into the taxonomic and functional diversity of microbial communities using minimally invasive sampling techniques.

In chapter 4, I include the first study wherein I present a comprehensive analysis of the lung metagenome of healthy individuals, including both smokers and nonsmokers, and diseased individuals with either cystic fibrosis or chronic obstructive pulmonary disease. I investigate the lung community composition for these health categories, detect the presence of antibiotic resistance genes, identify strain types of abundant community members and reconstruct near complete genome of a respiratory pathogen. An additional objective of this work was to assess the capability of sputum samples to provide information pertinent to the treatment of cystic fibrosis when processed using WGS sequencing. We summarize that WGS sequencing of CF sputum provides clinically relevant information in addition to current diagnostic methods and it can be an attractive alternate to some of the experimental methods which are labor-intensive and depend on an array of different technologies.

In chapter 5, I present the second study which characterizes the metagenome of calcified dental plaque in ancient individuals. The objective of this study was to profile the health and dietary habits of humans living approximately 1000 years ago. My contributions to this work include detailing the community composition of the calcified plaques, and providing evidence for the emergence of antibiotic resistance. I addition to this, I assembled a partial genome of an ancient strain of *Tanerella forsythia* which is a contemporary oral pathogen.

Besides the major contributions described above, I contributed to two other studies investigating *Salmonella typhimurium* infection in the mouse gut. The first study was aimed at understanding the role of microbiota-derived hydro-

gen during *Salmonella* infection. Here, I systematically mined the published gut metagenomes of 6 different species to identify genes encoding for hydrogen producing and consuming enzymes with the aim of assessing the availability of hydrogen in different guts. The second project was aimed at characterizing the dynamics of *Salmonella typhimurium* infection and the subsequent recovery episode. For this study, I analyzed changes in the expression pattern of a variety of genes over the course of infection with the objective of investigating the capacity of inflammatory markers to return to their baseline values following infection. The results of these two studies are briefly summarized in chapter 6.

3.2 Challenges

The analysis of WGS sequencing data is often complicated by non-conformity of derived data to the assumptions of standard bioinformatics tools, e.g. if the sample source is unconventional or a rarity. Such atypical samples are normally not considered during the design of mainstream analysis tools. This became problematic during the analysis of WGS sequencing data derived from the metagenome of ancient dental plaque. Due to the extreme age of the samples (approximately 1000 years), the extracted DNA was extremely fragmented. This lead to very short sequencing reads of unequal lengths ranging from 20 to 90 base pairs. Since most metagenomic assemblers are not designed to accommodate such samples, a variety of tools were tested and fine-tuned for the analysis of this unconventional data.

Part II

Results

Chapter 4

Sputum DNA sequencing in cystic fibrosis: non-invasive access to pathogen genome information and strain identity

Rounak Feigelman[1,2], Christian R. Kahlert[4,5], Florent Baty[3], Frank Rassouli[3], Rebekka L. Kleiner[3], Philip Kohler[3], Martin H. Brutsche[3], Christian von Mering[1,2]

[1] Institute of Molecular Life Sciences, University of Zurich, Switzerland
[2] Swiss Institute of Bioinformatics, Switzerland
[3] Pneumology and Sleep Medicine, Cantonal Hospital St. Gallen, Switzerland
[4] Division of Infectious Diseases and Hospital Epidemiology, Childrens Hospital of Eastern Switzerland, Switzerland
[5] Infectious Diseases and Hospital Epidemiology, Cantonal Hospital St. Gallen, Switzerland

Correspondence: mering@imls.uzh.ch
Emails: christian.kahlert@kssg.ch, florent.baty@kssg.ch, frank.rassouli@kssg.ch, rounak.vyas@imls.uzh.ch, martin.brutsche@kssg.ch, kohlerphilipp@hotmail.com
Submitted to: BMC microbiome, 28th July 2016

4.1 Abstract

Background
Cystic fibrosis (CF) is a life-threatening genetic disorder, characterized by chronic microbial lung infections due to abnormally viscous mucus secretions within airways. The clinical management of CF typically involves regular

respiratory-tract cultures in order to identify pathogens and to guide treatment. However, culture-based methods can miss atypical or slow-growing microbes. Furthermore, the isolated microbes are often not classified at the strain level due to limited taxonomic resolution.

Results

Here, we show that untargeted metagenomic sequencing of sputum DNA can provide valuable information beyond the possibilities of culture-based diagnosis. We sequenced the sputum of six CF patients and eleven control samples without prior depletion of human DNA or cell size selection, thus obtaining the most unbiased and comprehensive characterization of CF respiratory tract microbes to date. We present detailed descriptions of the CF and healthy lung microbiome, reconstruct near complete pathogen genomes, and confirm that CF lungs consistently exhibit reduced microbial diversity. Crucially, the obtained genomic sequences enabled a detailed identification of exact pathogen strain types, when analyzed in conjunction with existing multi-locus sequence typing databases. We also detected putative pathogenicity islands and indicators of antibiotics resistance, in good agreement with independent clinical tests.

Conclusions

Unbiased sputum metagenomics provides an in-depth profile of the lung pathogen microbiome, which is complementary to and more detailed than standard culture-based reporting. Furthermore, functional and taxonomic features of the dominant pathogens, including antibiotics resistances, can be deduced supporting accurate and non-invasive clinical diagnosis.

4.2 Background

Cystic fibrosis (CF) is one of the most prevalent genetic disorders in the Caucasian population, affecting about one in 2500 newborns [1]. This autosomal recessive condition affects mostly secretory organs, such as the pancreas, liver and lungs. CF is caused by mutations in the Cystic Fibrosis Transmembrane Regulator (CFTR) gene, whose protein product is involved in the transport of chloride ions across the apical membrane of epithelial and blood cells. Loss of CFTR protein function causes thickened extracellular mucus to accumulate, which impairs mucociliary clearance in the airways. CF prominently leads to microbial pathogen colonization in the lung, followed by recurrent pulmonary infection and chronic inflammation. Treatment options exist, including mechanical and enzymatic mobilization of mucus, drug therapy to improve residual CFTR function [2], antibiotics therapy to reduce pathogen load, anti-inflammatory drugs, and lung transplantations. Nevertheless, for the majority of patients the condition leads to progressive pulmonary damage, and eventually respiratory failure and death.

CF lungs are colonized by a number of pathogenic bacteria, commonly in-

cluding *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Haemophilus influenza* and *Burkholderia cepacia* [3]. While prompt and aggressive antibiotic therapies can often control infections, prolonged antibiotic treatments may favor the emergence of antibiotic resistances, and can facilitate colonization by multidrug resistant pathogens such as *Achromobacter xylosoxidans* and *Stenotrophomonas maltophilia* [4], [5]. Currently, culture-based techniques are routinely employed to identify and classify lung pathogens, often using selective culture media designed for specific groups of pathogens [6]. However, the culture conditions and procedures are necessarily biased towards known, previously encountered pathogens - whereas novel, slow-growing or rare microbes might potentially be missed (e.g. atypical mycobacteria). Meanwhile, the taxonomic identification of observed pathogens often has limited resolution, and the physiology and resistance profiles of the colonies are not backed up using genomic information. Lastly, the background community opportunistic or accidental members of the lung microbiome is not routinely studied for clinical use, despite its potential to harbor antibiotics resistance genes and to elicit or modulate immune responses.

Culture-independent, genomic sequencing techniques offer potential alternatives for identifying pathogens and opportunistic colonizers, and for guiding therapeutic decision-making. However, such methods are not routinely applied for CF management, with the exception of PCR-based surveys of the 16S ribosomal RNA gene [7], [8], which are of limited taxonomic and functional resolution. Here, we develop a pragmatic approach that aims to maximize molecular information, while minimizing patient discomfort and risk exposure. To achieve this, we sequence DNA from non-invasive sputum samples, without prior removal of host DNA and without complex enrichment- or depletion-protocols. Forgoing host DNA depletion yields a substantial fraction of sequence reads that are of human origin, and there will be contamination from the upper respiratory tract and mouth, but the simplicity of the approach has the unique advantage of providing an unbiased, comprehensive and reproducible set of reads from the deep lung as well. For the current study, we collected and sequenced sputum samples from the following categories: adult and pediatric CF patients, Chronic Obstructive Pulmonary Disease (COPD) patients, and healthy individuals. Using whole genome shotgun sequencing, we observed several known pulmonary pathogens whose genome coverage routinely exceeded 95%, allowing us to type the strains with very high precision using existing multi-locus sequence typing (MLST) databases. Patient-specific differences from reference strains are noted and discussed. Using the work flow we established, a detailed genomic profile can be generated for any dominant pathogen in the lung, including also potentially unknown pathogens.

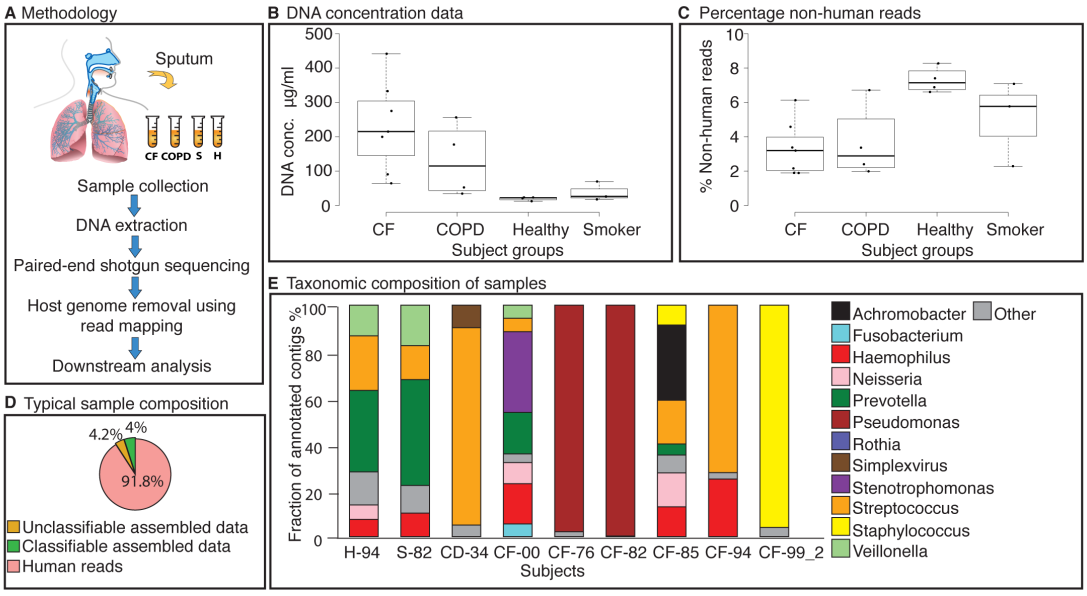

Figure 4.1: Sputum metagenomics workflow. a. Overview of the procedure. b. Concentration of extractable DNA in sputum, across subject groups. c. Fraction of non-human DNA sequence reads across subject groups. d. Fraction of DNA sequence reads of a representative healthy sample, further broken down according to taxonomic assignability to the assembled nucleotides from non-human fraction. e. Taxonomic composition of all taxonomically assignable, non-human sequences, at genus level (for each of the control groups, only one representative sample is shown). All genera constituting at least 4.5% of the annotated fraction in each sample are assigned a color code

4.3 Results and discussion

Sputum samples contain diagnostically useful DNA

We collected sputum samples from a total of 17 subjects: six CF patients, four patients with COPD, and seven healthy controls (three smokers, four non-smokers). We extracted the total DNA from each sputum sample and sequenced it without removal of human DNA, using the Illumina HiSeq 2000 platform, Figure 4.1A. In total, the average sequencing depth was around 30 million read pairs per sample (refer Figure 7.1). We then separated and quantified human and non-human read pairs; the latter were further assembled into contigs and annotated using homology-based searches (section 4.5). The observed DNA concentrations in sputum varied across subject groups, with CF and COPD patients presenting on average higher DNA concentrations than the healthy controls, Figure 4.1B. This observation agrees with previous studies which suggested that disintegrating inflammatory cells such as neutrophils release high concentrations of host DNA into CF and COPD sputum, leading to an increased sputum viscosity [9]. The observed higher DNA concentrations in

the disease groups were thus likely due to lung inflammation. Correspondingly, a large fraction of sequenced DNA ($>$90%) was of human origin, Figure 4.1C. Conversely, healthy subjects produced smaller volumes of sputum with lower DNA concentrations, and their residual DNA was also mostly of human origin (refer Figure 7.2).

Lung microbiota composition varies strongly across individuals

After limited (conservative) assembly of the non-human sequence fraction into contigs, approximately half of the assembled nucleotides could be assigned a taxonomic identity through homology searches, irrespective of subject groups (Figure 4.1D). For healthy subjects, we found an overall higher diversity (average Shannon entropy of 3.07 in healthy non-smoker lungs, versus 1.08 in the CF lungs, $p < 0.038$). Interestingly, the diversity was found to be reduced not only in the sputum of CF and COPD, but also to some extent in smokers (refer Figure 7.3). The most abundant taxa in healthy subjects were *Prevotella*, *Streptococcus*, *Veilonella*, *Haemophilus* and *Neisseria* (Figure 4.1E). Previous studies have indicated that the healthy lung does not harbor a stable and specific microbiome, but rather a mixture of microbes from the upper respiratory tract and oral cavity [10]-[13]. In agreement, many of the microbes we identified in healthy subjects corresponded to known oral (and occasionally also gastrointestinal or vaginal) flora.

In contrast, the microbiome composition in CF subjects was highly variable and distinct from standard oral microbiomes. Each patient harbored a unique and atypical community, often dominated by one or a few principle colonizers/pathogens such as *Pseudomonas*, *Staphylococcus*, *Stenotrophomonas*, or *Achromobacter*, (Figure 4.1E and Figure 7.4). Compositionally, the CF flora only marginally overlapped with those of the healthy and smoker populations.

Observed microbiota in CF subjects match clinical diagnosis

For five of the six CF subjects, all bacterial pathogens identified in the clinical culture diagnosis were also identified through the DNA sequencing, each with at least 10000 bp mapped to their genome (refer Table 7.1). For example, we confirmed the presence of the typical CF pathogens *Achromobacter* and *Staphylococcus* in the sample CF-85, despite ongoing antibiotic therapy. In this particular patient we also observed multiple instances of antibiotic resistance genes, including genes annotated to diminish effectiveness of the ongoing treatment (see below, Prediction of antibiotic resistances and fitness-conferring mutations). We also observed commensals such as *Prevotella, Neisseria, Streptococcus* and *Haemophilus* in lower abundances. Similarly, we detected *Pseu-*

domonas, a multidrug-resistant (MDR) *Stenotrophomonas,* and several other common microbial inhabitants in pediatric CF patient CF-00, despite ongoing antibiotic treatment. In contrast, the microbiomes of adult CF patients CF-82 and CF-76 were each primarily dominated by a single pathogen, *Pseudomonas,* with very low relative abundances of *Streptococcus*.

Patients CF-99 and CF-94 exhibited a microbial community largely dominated by *Staphylococcus* and *Streptococcus* respectively. Lack of a diverse microbial community was likely due to ongoing antibiotics treatment for patients CF-94 and CF-99 at the time of sample collection (Table 7.2 and Table 7.3 - 7.8); in this case, no resistances against the administered antibiotics had been detected in clinical testing. The use of antibiotics is known to correlate inversely with diversity and to instigate significant changes in bacterial community structure, especially in younger patients that harbor a relatively rich and susceptible microbial community [14], [15] as compared to older patients that often develop resilient communities [16], [17].

Samples from COPD patients reveal intermediate microbial complexity

We analyzed the microbiome diversity for each COPD patient and found that, of the four samples collected, CD-47 had the largest and most diverse population, closely resembling the composition of a healthy microbiome. Subject CD-34 was unique in that it was the sole sample in which we detected significant amounts of a virus, *Herpes simplex* virus. This virus was covered deeply enough to be partially assembled, and was seen against a background of *Streptococcus, Rothia* and *Haemophilus* with *Fusobacterium* and *Prevotella* greatly reduced (Figure 7.4). Overall, samples from COPD patients were somewhat more difficult to characterize. On several occasions we failed to detect eukaryotic genera that had been observed in the clinical culture-based diagnostics. For example, in patient CD-42 we were able to reliably detect the presence of several bacterial community members (confirmed by clinical microscopy results) but were unable to confirm the presence of *Candida*. Since eukaryotic genomes tend to be larger and more challenging to assemble, they may sometimes fail to be detected in sufficient numbers in our approach.

Entropy landscapes allow the detection of clonally expanded pathogens

The microbial community in the sputum of CF patients is expected to be heavily skewed a small number of entrenched, chronic pathogens stand out against a more diverse background of contaminants and putatively harmless colonizers [18], [19]. We devised a three-dimensional binning strategy adapted to this sit-

uation, in which each contig is assessed in terms of i) GC-content as a proxy for broad taxonomic identity, ii) length as a proxy for assembly depth, and iii) sequence homogeneity within the assembly as a proxy for clonality. The latter measure is expressed as entropy, where a small entropy value reflects a low number of mismatched sites in the assembly of a given contig. Low-entropy contigs should reflect clonal or near-clonal microbial strains (within the limits set by sequencing accuracy and depth). We used these three measures to visualize the entire non-human sequencing result of any patient of interest in a three-dimensional binning plot (Figure 4.2). For those contigs whose taxonomic identity could be confidently inferred, we additionally used a color code to highlight groups of sequences that might putatively belong to the same genus.

In Figure 4.2, entropy landscapes are used to visualize the lung community composition of two CF patients with distinct dominant pathogens, as well as one representative sample each from COPD, smoker and non-smoker groups. For CF-00, the plot shows two likely clonal overgrowths with distinct GC content (Figure 4.2A), suggesting chronic infections by two distinct pathogen species. Indeed, annotation revealed these contig groups to consist exclusively of members of the genera Stenotrophomonas and Haemophilus, respectively (Figure 4.2B).

These observed landscapes are characteristic of microbial communities with one or a few dominant members that have grown clonally to occupy a sizeable proportion of their niche. In contrast, the healthy and smoker groups generally were not dominated by few clonal species, as reflected in the absence of clustered low-entropy contigs (see Figure 7.5−7.8 for plots of each subject sampled).

The entropy landscapes allow the visual separation of likely oral contaminants and low-abundance colonizers from the clonal pathogen(s) growing chronically. Furthermore, any non-annotated contigs that visually cluster within the pathogen contigs may indicate undocumented genomic regions, which would have been recently introduced into the pathogen genome and may not be known from reference strains in databases. Hypothetically, even atypical pathogens that are not yet annotated in any database would become discernable, although we have not encountered such a case among our samples.

Multi locus sequence typing of pathogens using unbiased sputum sequences

Multi Locus Sequence Typing (MLST) is a well-established method to characterize isolates of a given microbial species in the context of previously observed strains of that species, via DNA sequencing of a limited, pre-defined group of diagnostic genes [20]. Traditional MLST requires the isolation and culture of microbes of interest, followed by specific PCR assays targeting the genes used

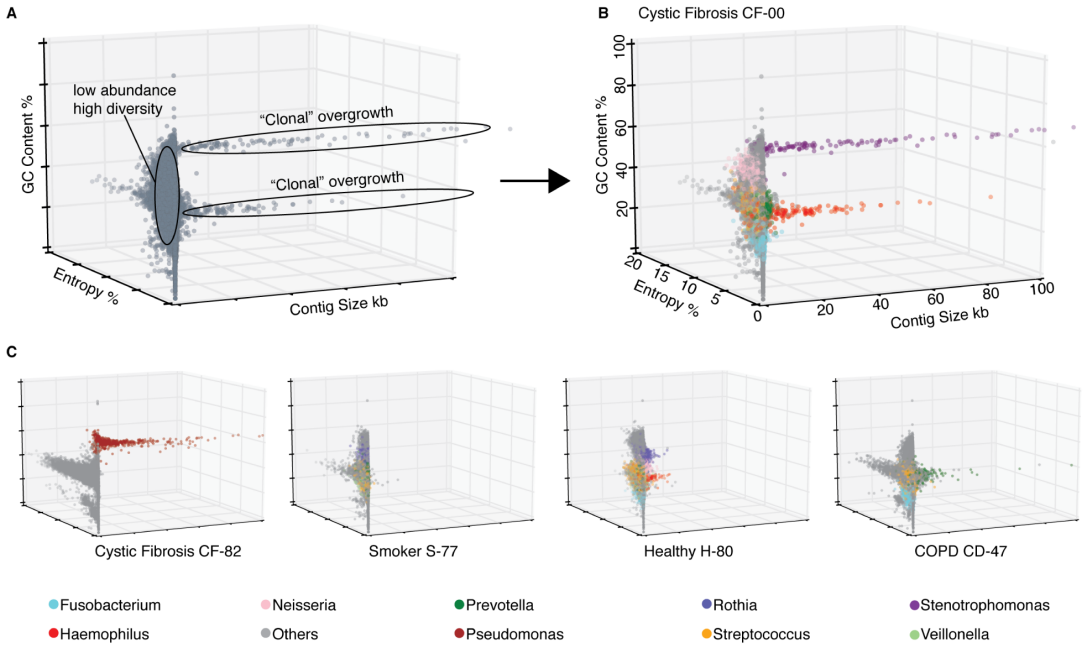

Figure 4.2: Pathogen overgrowth can be separated from background diversity. Sequence contig feature plots, depicting at least one sample from every subject group. Each data point represents an assembled contig, with colors corresponding to genus level taxonomy annotations. The three axes show contig length (X-axis), contig sequence heterogeneity (entropy, Y-axis) and GC-content (Z-axis). a. Magnified view of the plot of patient CF-00 without taxonomic annotation. b. same plot (CF-00), but with taxonomic annotation. c. Representative plots of one subject from each group. Throughout, genera constituting less than 5% of the annotated fraction, as well as un-annotated contigs, are shown in grey color.

for strain typing. In contrast, by using WGS sequencing data we omit these steps and directly proceed to characterizing strains of interest from the mixture. Importantly, there is no need to decide, ahead of the experiment, which strains are to be typed since no specific PCR is required. As long as a species or genus has been previously subjected to MLST genotyping (i.e., a well populated MLST database with corresponding marker genes is available), it can be characterized. In the following, we describe two examples of MLST strain genotyping in the sputum of CF patients one yielding a previously observed strain from a well-sampled strain collection, the other yielding a more exotic strain for which even the exact species designation remains unclear (it may belong to a new, as yet un-named species).

We characterized a strain of *Stenotrophomonas maltophilia* which we had observed in the sputum of patient CF-00. *S. maltophilia* is an intrinsically multi-drug-resistant (MDR) opportunistic pathogen that has been isolated from sev-

eral water-associated environments inside and outside of hospital premises [21]. Like many free-living opportunistic pathogens, it possesses a large and versatile genome that allows it to colonize diverse environments and degrade toxic compounds such as antibiotics, even using them as a food sources [22]. *S. maltophilia* exhibits high levels of genetic diversity, making it hard to precisely track the source of infections and distribution of isolates in hospitals. We performed MLST using a standard set of seven house-keeping genes [23], all of which were found with 100% sequence coverage in our sample. We observed that the patient harbored a single strain (likely from a single infection event), which was 100% identical to a strain encountered previously, in a CF case in the UK. We placed this strain, together with other strains observed previously, in a phylogenetic tree constructed from the MLST alignment, see Figure 4.3. The tree was then annotated with the sampling origin of each strain: clinical, environmental, hospital environment, or animal-associated. The phylogenetic analysis revealed a clear clustering of clinically obtained strains in a single clade, suggestive of specialization and frequent transfers from patient to patient. Overall, we found this genome to be very well recovered from the sputum, with an analysis using CheckM [24] reporting it to be 97.2% complete.

Next, we observed and characterized a putative *Achromobacter* strain from patient CF-85. Members of the *Achromobacter* genus form a group of gram negative, strictly aerobic, motile bacteria of which more than 10 species are currently known. The majority of strains isolated from CF patients belong to the species *A. xylosoxidans*, which is also an intrinsically multidrug-resistant opportunistic pathogen [25]. Outside of patients, strains can also be found in a variety of aquatic environments ranging from moist soils to dialysis solutions [26]. *Achromobacter* infections have been generally observed in older patients with pulmonary diseases, but their implication in deteriorating lung function has remained unclear [27], [28]. Accurate identification and discrimination of different *Achromobacter* species has been a challenging task due to limited taxonomic delineation. A recent study revealed that several commercial test systems used in different diagnostic laboratories were unable to distinguish different *Achromobacter* species infecting CF patients, and would often identify them incorrectly as *A. xylosoxidans* [29].

Using the appropriate gene set for this genus, we again performed MLST. In this case, we did not find any matches to previously documented strains. Instead, we placed our sequences on a phylogenetic tree encompassing the entire genus, constructed using concatenated housekeeping genes from all species and type strains available in the PUBMLST database (`http://www.pubmlst.org/smaltophilia/`) [30], see (Figure 4.4 and methods). Our observed strain did not cluster with other identified strains, with the exception of a single unnamed and uncharacterized isolate from another CF case. The closest neighbors of these two strains in the tree were *A. marplatensis* and *A. pulmo-*

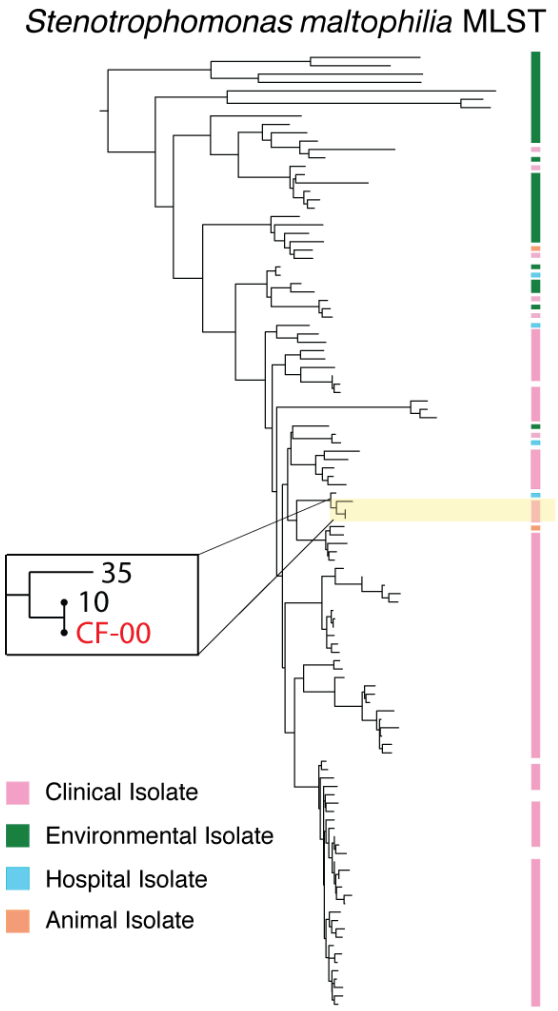

Figure 4.3: Strain typing of *S. maltophilia* strain from sputum sample of CF-00. Yellow color highlights the phylogenetic position of the strain observed in this study, relative to previously typed strains deposited in MLST databases. Isolation sources of database strains are shown color-coded.

nis. Although routine clinical analysis using microscopy had identified the strain as *A. xylosoxidans*, our phylogenetic analysis suggested otherwise; the two sequences were sufficiently removed from *A. xylosoxidans* to suggest a novel but previously unidentified clade. Independent studies [31] have also provided evidence to support the presence of as-yet unnamed and uncharacterized species in the *Achromobacter* genus, responsible for CF infections in patients. Further species divisions in this clinically relevant but under-sampled genus are needed, and patient-derived genomes such as ours might provide valuable context.

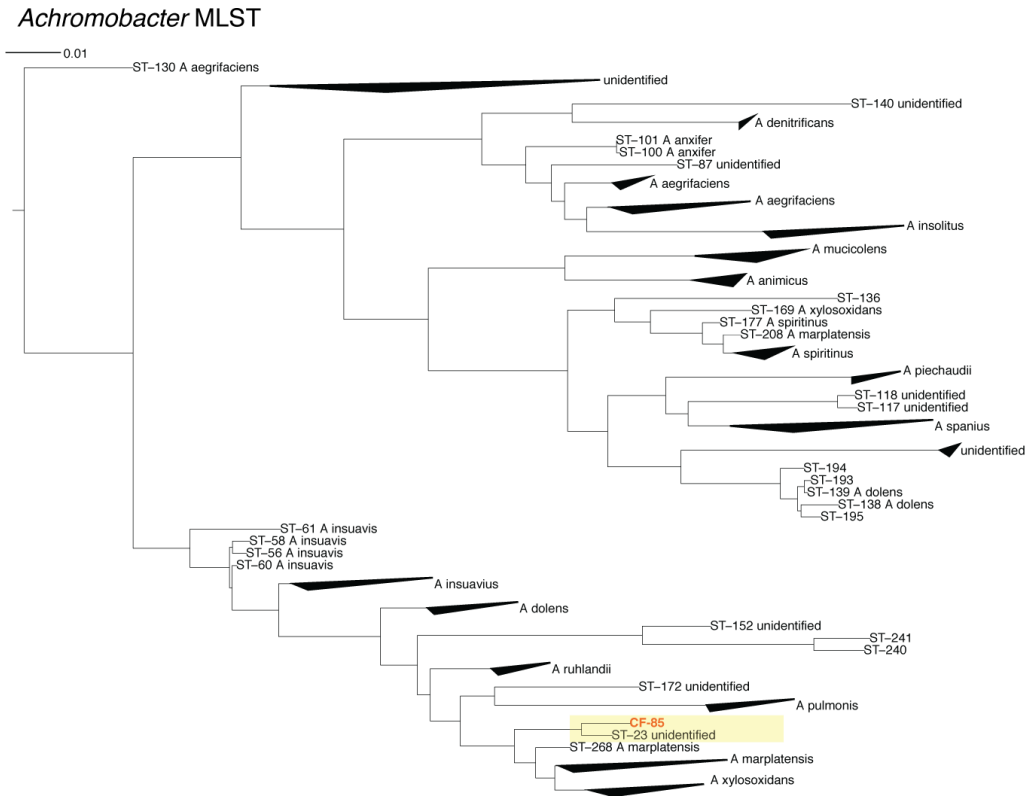

Figure 4.4: Strain typing for an *Achromobacter* isolate present in CF-85. No close relatives have been observed for this isolate (the strain likely does not belong to a named species). All monophyletic clades with 95% members from a single species have been collapsed.

Prediction of antibiotic resistances and fitness-conferring mutations

Chronic colonizers can adapt to their host environment to sustain themselves under varying selection pressures, such as antibiotic treatments or the presence of potentially competing co-infections. This is particularly problematic in the case of infections caused by antibiotic-resistant bacteria in CF patients. For example, a five-fold increase in MRSA infections has been observed in the past 15 years [32], and 18.1% of *P. aeruginosa* infections in a population of primarily young adult CF patients were reported to be MDR [32]. These pathogens acquire various resistance mechanisms to therapeutic agents including altered membrane permeability, efflux pumps or induced enzymatic modifications. Isolates with identical resistance patterns sometimes exhibit different genetic modifications, indicating that these bacteria can use distinct strategies to respond to similar environmental pressures [33].

To test whether sputum sequencing might help to guide antibiotic treatment, we screened for the presence of putative resistance-conferring genes and alleles. We compared all predicted open reading frames in each sample against a

database of known and annotated antibiotic resistance genes (see Methods), employing conservative similarity cutoffs. For example, in the sputum of CF-85, we identified genes encoding for class A beta lactamases, which are classified as serine enzymes conferring resistance to penicillin. In the same sample, we also identified a 23S ribosomal RNA methyltransferase, conferring varying degrees of resistance to macrolide, lincosamide and streptogramin b antibiotics, see (Table 4.1 and Table 7.3 - 7.8).

Tested Antibiotics	Clincal Results	Genomic Prediction
Penicillin G	Resistant	Resistant
Ampicillin	Resistant	Not detected
Oxacillin	Sensitive	Not detected
Amoxicillin-Calv	Sensitive	Not detected
Cefazolin	Sensitive	Not detected
Cefamandol	Sensitive	Not detected
cefuroxim	Sensitive	Not detected
Imipenem	Sensitive	Not detected
Meropenem	Sensitive	Not detected
Azithromycin	Resistant	Resistant
Clarithromycin	Resistant	Resistant
Erythromycin	Resistant	Resistant
Tetracyclin	Sensitive	Not detected
Ciprofloxacin	Resistant	Not detected
Co-Trimoxazol	Sensitive	Not detected
Rifampicin	Sensitive	Not detected
Clindamycin	Resistant	Resistant
Fusidinsaure	Sensitive	Not detected

Table 4.1: The *Achromobacter* strain isolated from patient CF-85 underwent routine clinical testing for antibiotics sensitivity; the compounds tested and the observed results are shown. This is contrasted with automated predictions based on the gene content of the sputum sequence data.

The predictions were validated independently by clinical resistance reports, where the majority of both resistant as well as non-resistant calls were confirmed (Table 4.2). False predictions were limited to false negatives, i.e., clinically observed resistances which were not predicted based on sequence analysis. Interestingly, we additionally observed a virulence gene, mprF, annotated as providing resistance against naturally secreted antimicrobial peptides known as defensins. These cationic peptides are largely secreted by neutrophils and by the airway epithelium in the CF lungs to protect the epithelia against infections. *S. aureus* strains with resistance to defensins show a greater pathogenic potential [34]. Thus, the presence of such virulence factors is of general relevance to clinicians when designing treatments for CF patients.

Apart from antibiotics resistances, other phenotypes such as biofilm formation or exo-polysaccharide secretion may also be actively adapting in chronic

Subject ID	True Positive	True Negative	False Positve	False Negative
CF-85	5	11	0	2
CF-82	1	9	0	1
CF-76	0	11	0	0
CF-00	0	7	0	0
CF-94	0	16	0	2
CF-99	0	17	0	3
CF-99_2	0	16	0	2

Table 4.2: Prediction of antibiotics resistances. b. Summary table for all CF subjects, indicating the overlap between the resistance predictions and the clinical test results.

colonizers. Indeed, it is known that chronic colonizers develop considerable genomic heterogeneity which is perhaps maintained by specialization or balanced selection [35], [36]. Given the depth of sequencing used here, it is not possible to study this heterogeneity quantitatively, but specific mutations of interest can be tracked. For example, initial isolates of *Pseudomonas aeruginosa* from CF patients are generally non-mucoid and responsive to antibiotics. However, during protracted infections, these pathogens start overproducing an exopolysaccharide known as alginate which is a polymer of -D-manuronic acid and L-glucuronic acid [37], eventually leading to their transition into a mucoid phenotype. Mucoid *P. aeruginosa* are immune to several antibiotics and to phagocytosis [38]. Correspondingly, the mucoid phenotype is directly linked with poor clinical outcome for patients. According to the clinical laboratory report, the sputum of CF-82 harbored both mucoid and non-mucoid *P. aeruginosa* strains. To better understand the underlying genetic modifications that lead to this phenotypic transition, we inspected the mucA gene, which encodes for a transmembrane -factor responsible for limiting expression of the 12 gene alginate operon (algA-algD). Loss of function mutations in the mucA gene typically result in production of alginate, in turn giving rise to a mucoid phenotype. In CF-82, we indeed identified 11 sequence reads showing a single-nucleotide deletion at position 429 in the mucA gene (Figure 4.5), leading to a truncated and presumably non-functional protein. In contrast, 7 reads supported the presence of a non-mutated, fully functional protein. Since there was little heterogeneity elsewhere in the genome, this is indeed suggestive of an ongoing adaptation.

Genome comparisons reveal patient-specific pathogen features

The MLST procedure allowed us to precisely characterize the taxonomic identity of strains of interest, but provided little phenotypic information regarding pathogenicity or metabolic characteristics. Moreover, this information is not routinely available from culture-independent techniques in the clinic. To address

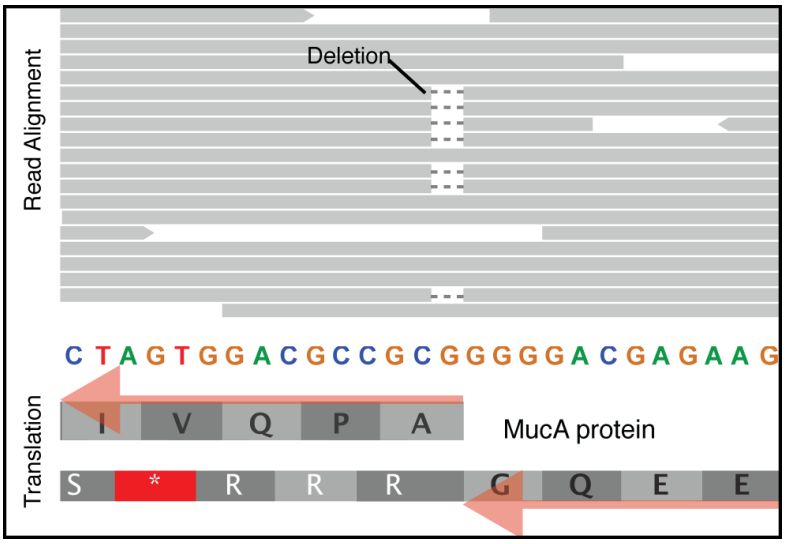

Figure 4.5: Read alignment against a section of the mucA gene from *Pseudomonas aeruginosa*, from patient CF-82. 11 reads show a wild-type sequence at this position, but 7 reads show a deletion event predicting a non-functional protein and a corresponding shift from a non-mucoid to a mucoid phenotype in this strain.

this, we collected all contigs assembled from CF-00 belonging to *S. maltophilia*, and aligned them against two closely related clinical database strains, Sm K279a and Sm ISSMS3 [genbank accession NC_010943 and NZ_CP011010, respectively] (Figure 7.9). This identified seven large-scale homologous regions, see (Figure 4.6A). These regions were interspersed by non-homologous intervals unique to each genome, including some in the assembled genome from patient CF-00. The assembled genome also exhibited some genomic rearrangements with respect to the reference strains. We next made gene predictions for all the unique genomic regions of size 1kb or greater present in the reconstructed genome. Interestingly, we found a large region of 23 genes (Figure 4.6B), 14 of which were found to code for a virulence-associated Type VI Secretion System (T6SS). T6SS systems typically consist of a conserved cluster of 13 core genes, 10 of which were observed in our gene set together with 4 additional non-conserved T6SS accessory genes (Figure 4.6C) responsible for post-translational regulation based on orthology predictions. T6SS secretion systems (Figure 4.6D) were first described a decade ago in *Vibrio cholerae* [39] and since then have been studied in several gram-negative bacteria including *P. aeruginosa* [40]. They allow the secretion of a range of substrates such as toxins, adhesins, hydrolytic enzymes and effector proteins, and have been classified into four subtypes [41]. T6SS have been associated with biofilm formation, and antagonistic or bactericidal functions towards competing bacterial species [40]. They are frequently observed in genomic islands and their presence shows little correlation to bacterial taxonomy, suggesting that they are

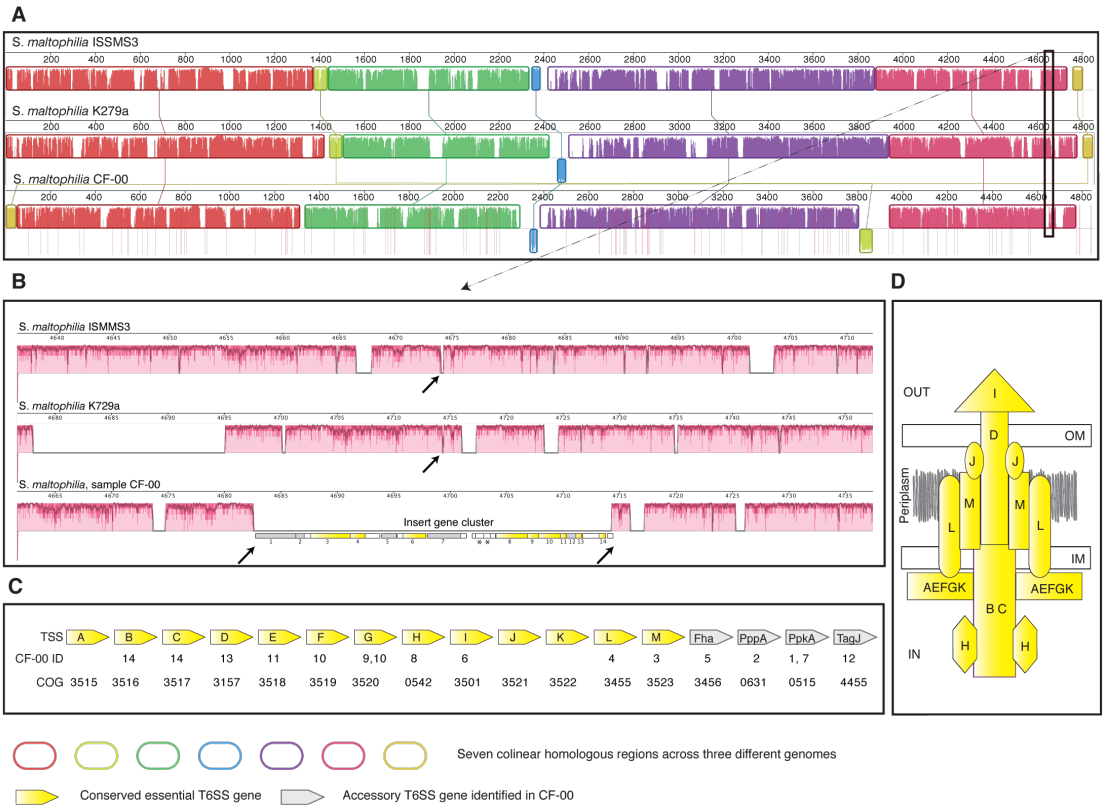

Figure 4.6: Pathogen genome comparisons reveal patient-specific additions. Two public reference genomes of *S. maltophilia* are compared against assembled contigs from patient CF-00. a. Genome wide alignment showing blocks of colinearity, additions and deletions. White stretches indicate un-alignable, unique regions in each genome. Vertical red lines separate individual assembled contigs. b. Magnified view centered on a genomic region that is unique to the strain in patient CF-00. Genes with homology to Type Six Secretion Systems (T6SS) have been labeled with numerical IDs (see panel C below). Genes marked with an Asterisk showed no detectable homology in sequence databases. c. The core gene cluster of T6SS is depicted in yellow; additional accessory T6SS genes which are also observed in patient CF-00 are colored in grey. d. Schematic model of the T6SS protein structure based on present knowledge.

frequently acquired through horizontal gene transfer [42].

In the past decade, T6SS have repeatedly been reported to be absent in all *S. maltophilia* isolates, both in clinical and environmental isolates [43]-[45]. Its presence in our clinical isolate thus highlights its outstanding ability to adapt to new hosts and surroundings. In addition to our observation, recent observations in new *S. maltophilia* strains [46] were also supportive of T6SS systems (although the genomes are yet to be published).

The spread of such virulent secretory systems in multidrug-resistant CF pathogens provides a unique opportunity to identify new targets for designing

innovative treatments. Anti-virulence drugs [47] that target specific secretion systems to disarm the bacteria can be a likely alternative to conventional antibiotics. Thus, whole genome reconstruction of highly abundant bacteria in the patient samples provides an exclusive possibility to study idiosyncratic genomic regions and to identify potential drug targets for targeted treatments.

4.4 Conclusions

In this study, we have introduced a culture-independent technique for characterizing airway pathogens in the chronically inflamed lung, using routine non-invasive sputum sampling coupled to unbiased WGS sequencing. Our approach provides additional valuable information complementing routine culture-based clinical microbiology results in the future. This could help to design tailored treatment regimes by reducing the risk of ineffective treatment.

We find, firstly, that WGS sequencing can serve to describe the broad taxonomic composition of the lung microbiome, particularly when combined with reference databases. Reference information for human-associated microbes is growing at a remarkable pace [48], [49] and should make WGS-based taxonomic classification ever more accurate in the future.

Secondly, various binning approaches [50] can be used to partially assemble genomes of interest from the WGS data. In the case of chronic infections originating from one or a few clonal invasions, we find that contig-by-contig entropy is a good measure for isolating pathogens from contaminants and the complex colonizing background of microbes. Such an approach may help clinicians to more precisely identify the disease-causing pathogen.

Thirdly, for those pathogens that have already been well studied in molecular terms, we demonstrated that MLST can be applied directly using WGS data. This allows the exact strain identity to be established, leveraging the power of MLST databases and catalogued strain observations. Importantly, this can significantly simplify the process of tracing infection outbreaks at clinics using untargeted, retrospective data.

Lastly, the partially assembled genomes and the remaining, unassembled contigs can be informative with regard to the expected efficacy of treatment options. We find that while antibiotics resistance prediction are substantially supported by confirmatory clinical tests, they are not yet entirely error free. This is likely to improve with better data curation in the resistance databases, but the most challenging resistances to predict correctly will be those that arise from specific mutations in normal, cellular genes [51]. Importantly, aside from predicting resistances, WGS data may guide the decision as to which antibiotics to include in clinical testing in the first place, particularly for second-generation antibiotics designed to counteract or circumvent known resistances.

Overall, WGS sequencing of sputum may become one of the building blocks supporting the advancement of a more personalized medicine. It yields not only deep insights into the lung microbiome by allowing an unbiased metagenomic dissection of microbial pathogens, but also enables analysis of human genomic DNA for host genotyping (e.g. for host susceptibility to infection or for unexpected treatment responses). For routine tracking, deep WGS sequencing could be alternated with more shallow survey sequencing or 16S sequencing; the latter are likely sufficient to quantify changes in community composition, with deep sequencing only necessary when new pathogens invade.

4.5 Methods

Sputum samples and DNA extraction

Sputum was either produced spontaneously (in the case of CF and COPD patients), or after induction by hypertonic saline nebulization (in the case of healthy control subjects). The sampling was conducted at the Cantonal Hospital St. Gallen and at the Childrens Hospital of Eastern Switzerland. Healthy control subjects were free of symptoms of respiratory discomfort and did not show overt infections. All study participants provided informed consent. The study was approved by the cantonal ethical committee St. Gallen (EKSG $13/112$). The sputum samples were weighed and aliquoted into sterile tubes. After dilution in Sputolysin (Calbiochem Corp, San Diego, CA, USA), total DNA was extracted using the High Pure PCR template preparation kit (Roche, Basel, Switzerland) according to the manufacturers instructions. DNA concentration was measured using the ACTgene UV99 spectro-photometer at a wavelength of 260nm, and samples were stored at 20°C.

Whole genome shotgun sequencing

The TruSeq DNA Sample Prep Kit v2 (Illumina Inc, California, USA) was used for library generation. The quality and quantity of the enriched libraries were validated using a Qubit (1.0) fluorometer and the Caliper GX LabChip GX (Caliper Life Sciences, Inc., USA). The libraries were normalized to 10nM in Tris-Cl 10 mM, pH8.5 with 0.1% Tween 20. The TruSeq PE Cluster Kit v3-cBot-HS (Illumina, Inc, California, USA) was used for cluster generation using 2 pM of pooled normalized libraries on the cBOT. Paired-end sequencing was performed on the Illumina HiSeq 2000 at 2 x 101 bp using the TruSeq SBS Kit v3-HS (Illumina, Inc, California, USA). Reads were quality-checked with FastQC [52].

Removal of the host genome and assembly

We used Bowtie2 [53] to align the paired-end reads against the human reference genome, assembly Hg19 [54]. Read pairs were omitted from any further downstream analysis if one or both mates from the pair aligned to the human genome. We assembled contigs from the remaining read pairs using the Spades assembler [55] under the only assembly setting.

Taxonomic annotation and diversity estimation

We searched the contigs against the NCBI nucleotide database using Blastn [56] with an e-value cutoff of e^{-15}. The most common recent ancestor of all genomic sequences that aligned to a given contig with a bit-score within ten percent range of the highest scoring alignment was used to taxonomically annotate the contig. We discarded all contigs with annotations belonging to the metazoan kingdom, to further remove host genome sequences from further downstream analysis. We used nucleotide counts from assembled contigs with genus level taxonomic assignments to calculate Shannon entropy as a measure of diversity. We used Mann Whitney U tests for significance testing and subsequently adjusted the p-values using the Bonferroni correction for multiple testing.

Contig binning via entropy landscapes

We recruited all paired-end reads that had contributed to the assembly process back against the non-human contigs using Bowtie2. This recruitment was used to calculate the average number of mismatches and gaps over the length of the contig (entropy). This score was depicted on the z-axis of the microbial landscape plots, together with contig length on the x-axis, and GC content on the y-axis. Contigs from genera constituting more than 5% of the annotated non-human contigs were color-coded according to their annotation. We depicted the un-annotated and the remaining contigs from the low abundance genera in a single color.

MLST and phylogenetic placement of abundant clonal species

We constructed a maximum likelihood tree for *Stenotrophomonas maltophilia* using RaxML [57] under the GTRCAT model, with 1000 bootstraps, using the concatenated sequence composed of the seven house-keeping genes atpD, gapA, guaA, mutM, nuoD, ppsA and recA. In patient CF-00, these genes showed 100% identity over their entire length to a previously observed strain. All the sequences of the previously typed strains used for building the tree were downloaded from http://pubmlst.org/smaltophilia/. The tree was rooted using

Xanthomonas Campestris 8004, [genbank accession NC_007086.1] as an out-group. For the *Achromobacter* genus, we built another phylogenetic tree using the concatenated sequences of the seven house-keeping genes eno, gltB, lepA, nrdA, nuoL, nusA and rpoB. Full-length sequences of all seven genes were recovered from the assembled contigs of the patient CF-85. There was no exact match found to any previously documented strain. The sequences of previously typed strains used for building the tree were downloaded from `http://pubmlst.org/achromobacter`. The tree was rooted using *Bordetella pertussis* [genbank accession LN849008] as an out-group.

Identification of antibiotic resistance genes and cassettes

We used the comprehensive antibiotic resistance database [58] to search for resistance-conferring genes in our samples. The results from the web server were filtered to include only matches with over 90 amino acid length and over 90% identity. We further compared these findings with the clinical laboratory reports on observed antibiotics resistances.

Genome alignment

We used Mauve [59] for ordering the assembled *Stenotrophomonas* contigs from CF-00 and performing whole genome alignments against reference strains.

Bibliography

[1] S. Walters and A. Methta, Epidemiology of cystic fibrosis, in Cystic Fibrosis, Third Edition, CRC Press, 2012, pp. 16.

[2] S. C. Bell, K. De Boeck, and M. D. Amaral, New pharmacological approaches for cystic fibrosis: promises, progress, pitfalls., Pharmacol. Ther., vol. 145, pp. 1934, Jan. 2015.

[3] G. A. O. Laura M Filkins, Cystic Fibrosis Lung Infections: Polymicrobial, Complex, and Hard to Treat, pp. 18, Dec. 2015.

[4] K. Tan, S. P. Conway, K. G. Brownlee, C. Etherington, and D. G. Peckham, Alcaligenes infection in cystic fibrosis, Pediatr Pulmonol., vol. 34, no. 2, pp. 101104, Jul. 2002.

[5] I. Talmaciu, L. Varlotta, J. Mortensen, and D. V. Schidlow, Risk factors for emergence of Stenotrophomonas maltophilia in cystic fibrosis., Pediatr Pulmonol., vol. 30, no. 1, pp. 1015, Jul. 2000.

[6] J. Zhou, E. Garber, M. Desai, and L. Saiman, Compliance of clinical microbiology laboratories in the United States with current recommendations for processing respiratory tract specimens from patients with cystic fibrosis., Journal of Clinical Microbiology, vol. 44, no. 4, pp. 15471549, Apr. 2006.

[7] R. P. Dickson, J. R. Erb-Downward, F. J. Martinez, and G. B. Huffnagle, The Microbiome and the Respiratory Tract., Annu. Rev. Physiol., vol. 78, pp. 481504, Feb. 2016.

[8] B. Coburn, P. W. Wang, J. Diaz Caballero, S. T. Clark, V. Brahma, S. Donaldson, Y. Zhang, A. Surendra, Y. Gong, D. Elizabeth Tullis, Y. C. W. Yau, V. J. Waters, D. M. Hwang, and D. S. Guttman, Lung microbiota across age and disease stage in cystic fibrosis., Sci Rep, vol. 5, p. 10241, 2015.

[9] A. L. Smith, G. Redding, C. Doershuk, D. Goldmann, E. Gore, B. Hilman, M. Marks, R. Moss, B. Ramsey, T. Roblo, R. H. Schwartz, M. J. Thomassen, J. Williams-Warren, A. Weber, R. W. Wilmott, H. D. Wilson, and R. Yogev, Sputum changes associated with therapy for endobronchial exacerbation in cystic fibrosis, The Journal of Pediatrics, vol. 112, no. 4, pp. 547554, Apr. 1988.

[10] E. S. Charlson, K. Bittinger, J. Chen, J. M. Diamond, H. Li, R. G. Collman, and F. D. Bushman, Assessing Bacterial Populations in the Lung by Replicate Analysis of Samples from the Upper and Lower Respiratory Tracts, PLoS ONE, vol. 7, no. 9, p. e42786, Sep. 2012.

[11] E. S. Charlson, K. Bittinger, A. R. Haas, A. S. Fitzgerald, I. Frank, A. Yadav, F. D. Bushman, and R. G. Collman, Topographical Continuity of Bacterial Populations in the Healthy Human Respiratory Tract, American Journal of Respiratory and Critical Care Medicine, vol. 184, no. 8, pp. 957963, Oct. 2011.

[12] C. M. Bassis, J. R. Erb-Downward, R. P. Dickson, C. M. Freeman, T. M. Schmidt, V. B. Young, J. M. Beck, J. L. Curtis, and G. B. Huffnagle, Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals., MBio, vol. 6, no. 2, p. e00037, 2015.

[13] A. Venkataraman, C. M. Bassis, J. M. Beck, V. B. Young, J. L. Curtis, G. B. Huffnagle, and T. M. Schmidt, Application of a neutral community model to assess structuring of the human lung microbiome., MBio, vol. 6, no. 1, 2015.

[14] J. Zhao, P. D. Schloss, L. M. Kalikin, L. A. Carmody, B. K. Foster, J. F. Petrosino, J. D. Cavalcoli, D. R. VanDevanter, S. Murray, J. Z. Li, V. B.

Young, and J. J. LiPuma, Decade-long bacterial community dynamics in cystic fibrosis airways., Proc. Natl. Acad. Sci. U.S.A., vol. 109, no. 15, pp. 58095814, Apr. 2012.

[15] L. A. Carmody, J. Zhao, P. D. Schloss, J. F. Petrosino, S. Murray, V. B. Young, J. Z. Li, and J. J. LiPuma, Changes in cystic fibrosis airway microbiota at pulmonary exacerbation., Ann Am Thorac Soc, vol. 10, no. 3, pp. 179187, Jun. 2013.

[16] P. S. Brown, C. E. Pope, R. L. Marsh, X. Qin, S. McNamara, R. Gibson, J. L. Burns, G. Deutsch, and L. R. Hoffman, Directly sampling the lung of a young child with cystic fibrosis reveals diverse microbiota., Ann Am Thorac Soc, vol. 11, no. 7, pp. 10491055, Sep. 2014.

[17] A. A. Fodor, E. R. Klem, D. F. Gilpin, J. S. Elborn, R. C. Boucher, M. M. Tunney, and M. C. Wolfgang, The adult cystic fibrosis airway microbiota is stable over time and infection type, and highly resilient to antibiotic treatment of exacerbations., PLoS ONE, vol. 7, no. 9, p. e45001, 2012.

[18] N. Cramer, L. Wiehlmann, and B. Tmmler, Clonal epidemiology of Pseudomonas aeruginosa in cystic fibrosis., Int. J. Med. Microbiol., vol. 300, no. 8, pp. 526533, Dec. 2010.

[19] C. P. Coutinho, S. C. Dos Santos, A. Madeira, N. P. Mira, A. S. Moreira, and I. S-Correia, Long-term colonization of the cystic fibrosis lung by Burkholderia cepacia complex bacteria: epidemiology, clonal variation, and genome-wide expression alterations., Front Cell Infect Microbiol, vol. 1, p. 12, 2011.

[20] M. C. J. Maiden, M. J. Jansen van Rensburg, J. E. Bray, S. G. Earle, S. A. Ford, K. A. Jolley, and N. D. McCarthy, MLST revisited: the gene-by-gene approach to bacterial genomics., Nat. Rev. Microbiol., vol. 11, no. 10, pp. 728736, Oct. 2013.

[21] J. S. Brooke, Stenotrophomonas maltophilia: an Emerging Global Opportunistic Pathogen, Clinical Microbiology Reviews, vol. 25, no. 1, pp. 241, Jan. 2012.

[22] J. L. Martnez, Antibiotics and antibiotic resistance genes in natural environments, Science, 2008.

[23] G. Gherardi, R. Creti, A. Pompilio, and G. Di Bonaventura, Diagnostic Microbiology and Infectious Disease, Diagnostic Microbiology and Infectious Disease, vol. 81, no. 3, pp. 219226, Mar. 2015.

[24] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes., Genome Res., vol. 25, no. 7, pp. 10431055, Jul. 2015.

[25] C. R. Hansen, T. Pressler, K. G. Nielsen, P. . Jensen, T. Bjarnsholt, and N. Hiby, Inflammation in Achromobacter xylosoxidans infected cystic fibrosis patients., J. Cyst. Fibros., vol. 9, no. 1, pp. 5158, Jan. 2010.

[26] L. Amoureux, J. Bador, S. Fardeheb, C. Mabille, C. Couchot, C. Massip, A. L. Salignon, G. Berlie, V. Varin, and C. Neuwirth, Detection of Achromobacter xylosoxidans in Hospital, Domestic, and Outdoor Environmental Samples and Comparison with Human Clinical Isolates, Applied and Environmental Microbiology, vol. 79, no. 23, pp. 71427149, Nov. 2013.

[27] F. De Baets, P. Schelstraete, S. Van Daele, F. Haerynck, and M. Vaneechoutte, Achromobacter xylosoxidans in cystic fibrosis: Prevalence and clinical relevance, Journal of Cystic Fibrosis, vol. 6, no. 1, pp. 7578, Jan. 2007.

[28] C. Rnne Hansen, T. Pressler, N. Hiby, and M. Gormsen, Chronic infection with Achromobacter xylosoxidans in cystic fibrosis patients; a retrospective case control study, Journal of Cystic Fibrosis, vol. 5, no. 4, pp. 245251, Dec. 2006.

[29] T. Spilker, P. Vandamme, and J. J. LiPuma, Identification and distribution of Achromobacter species in cystic fibrosis, Journal of Cystic Fibrosis, vol. 12, no. 3, pp. 298301, May 2013.

[30] K. A. Jolley and M. C. Maiden, BIGSdb: Scalable analysis of bacterial genome variation at the population level, BMC Bioinformatics, vol. 11, no. 1, p. 595, 2010.

[31] W. Ridderberg, M. Wang, and N. Norskov-Lauritsen, Multilocus Sequence Analysis of Isolates of Achromobacter from Patients with Cystic Fibrosis Reveals Infecting Species Other than Achromobacter xylosoxidans, Journal of Clinical Microbiology, vol. 50, no. 8, pp. 26882694, Jul. 2012.

[32] Cystic Fibrosis Foundation, Cystic Fibrosis Patient Registry 2014 Annual Data Report, pp. 192, Oct. 2015.

[33] L. Yang, L. Jelsbak, R. L. Marvig, S. Damkir, C. T. Workman, M. H. Rau, S. K. Hansen, A. Folkesson, H. K. Johansen, O. Ciofu, N. Hiby, M. O. A. Sommer, and S. Molin, Evolutionary dynamics of bacteria in a human host environment, 2011.

[34] A. Peschel, R. W. Jack, M. Otto, L. V. Collins, P. Staubitz, G. Nicholson, H. Kalbacher, W. F. Nieuwenhuizen, G. Jung, A. Tarkowski, K. P. van Kessel, and J. A. van Strijp, Staphylococcus aureus resistance to human defensins and evasion of neutrophil killing via the novel virulence factor MprF is based on modification of membrane lipids with l-lysine., J. Exp. Med., vol. 193, no. 9, pp. 10671076, May 2001.

[35] T. D. Lieberman, K. B. Flett, I. Yelin, T. R. Martin, A. J. McAdam, G. P. Priebe, and R. Kishony, Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures., Nat. Genet., vol. 46, no. 1, pp. 8287, Jan. 2014.

[36] J. Diaz Caballero, S. T. Clark, B. Coburn, Y. Zhang, P. W. Wang, S. L. Donaldson, D. E. Tullis, Y. C. W. Yau, V. J. Waters, D. M. Hwang, and D. S. Guttman, Selective Sweeps and Parallel Pathoadaptation Drive Pseudomonas aeruginosa Evolution in the Cystic Fibrosis Lung., MBio, vol. 6, no. 5, pp. e0098115, 2015.

[37] G. Pulcrano, D. V. Iula, V. Raia, F. Rossano, and M. R. Catania, Different mutations in mucA gene of Pseudomonas aeruginosa mucoid strains in cystic fibrosis patients and their effect on algU gene expression., New Microbiol., vol. 35, no. 3, pp. 295305, Jul. 2012.

[38] Z. Li, M. R. Kosorok, P. M. Farrell, A. Laxova, S. E. H. West, C. G. Green, J. Collins, M. J. Rock, and M. L. Splaingard, Longitudinal development of mucoid Pseudomonas aeruginosa infection and lung disease progression in children with cystic fibrosis., JAMA, vol. 293, no. 5, pp. 581588, Feb. 2005.

[39] S. Pukatzki, A. T. Ma, A. T. Revel, D. Sturtevant, and J. J. Mekalanos, Type VI secretion system translocates a phage tail spike-like protein into target cells where it cross-links actin., Proc. Natl. Acad. Sci. U.S.A., vol. 104, no. 39, pp. 1550815513, Sep. 2007.

[40] R. D. Hood, P. Singh, F. Hsu, T. Gvener, M. A. Carl, R. R. S. Trinidad, J. M. Silverman, B. B. Ohlson, K. G. Hicks, R. L. Plemel, M. Li, S. Schwarz, W. Y. Wang, A. J. Merz, D. R. Goodlett, and J. D. Mougous, A type VI secretion system of Pseudomonas aeruginosa targets a toxin to bacteria., Cell Host and Microbe, vol. 7, no. 1, pp. 2537, Jan. 2010.

[41] L. E. Bingle, C. M. Bailey, and M. J. Pallen, Type VI secretion: a beginner's guide, Current Opinion in Microbiology, vol. 11, no. 1, pp. 38, Feb. 2008.

[42] F. Boyer, G. Fichant, J. Berthod, Y. Vandenbrouck, and I. Attree, Dissecting the bacterial type VI secretion system by a genome wide in silico analysis:

what can be learned from available microbial genomic resources?, BMC Genomics, vol. 10, no. 1, pp. 114, 2009.

[43] R. P. Ryan, S. Monchy, M. Cardinale, S. Taghavi, L. Crossman, M. B. Avison, G. Berg, D. van der Lelie, and J. M. Dow, The versatility and adaptation of bacteria from the genus Stenotrophomonas, pp. 112, Jul. 2009.

[44] K. L. Ormerod, N. M. George, J. A. Fraser, C. Wainwright, and P. Hugenholtz, Comparative genomics of non-pseudomonal bacterial species colonising paediatric cystic fibrosis patients, PeerJ, vol. 3, no. 23, p. e1223, 2015.

[45] P. Alavi, M. R. Starcher, G. G. Thallinger, C. Zachow, H. M. Iler, and G. Berg, Stenotrophomonas comparative genomics reveals genes and functions that differentiate beneficial and pathogenic bacteria, vol. 15, no. 1, pp. 115, Jun. 2014.

[46] M. Adamek, B. Linke, and T. Schwartz, Virulence genes in clinical and environmental Stenotrophomas maltophilia isolates: A genome sequencing and gene expression approach, Microbial Pathogenesis, vol. 67, no. C, pp. 2030, Feb. 2014.

[47] C. Baron, Antivirulence drugs to target bacterial secretion systems., Current Opinion in Microbiology, vol. 13, no. 1, pp. 100105, Feb. 2010.

[48] Human Microbiome Jumpstart Reference Strains Consortium, K. E. Nelson, G. M. Weinstock, S. K. Highlander, K. C. Worley, H. H. Creasy, J. R. Wortman, D. B. Rusch, M. Mitreva, E. Sodergren, A. T. Chinwalla, M. Feldgarden, D. Gevers, B. J. Haas, R. Madupu, D. V. Ward, B. W. Birren, R. A. Gibbs, B. Meth, J. F. Petrosino, R. L. Strausberg, G. G. Sutton, O. R. White, R. K. Wilson, S. Durkin, M. G. Giglio, S. Gujja, C. Howarth, C. D. Kodira, N. Kyrpides, T. Mehta, D. M. Muzny, M. Pearson, K. Pepin, A. Pati, X. Qin, C. Yandava, Q. Zeng, L. Zhang, A. M. Berlin, L. Chen, T. A. Hepburn, J. Johnson, J. McCorrison, J. Miller, P. Minx, C. Nusbaum, C. Russ, S. M. Sykes, C. M. Tomlinson, S. Young, W. C. Warren, J. Badger, J. Crabtree, V. M. Markowitz, J. Orvis, A. Cree, S. Ferriera, L. L. Fulton, R. S. Fulton, M. Gillis, L. D. Hemphill, V. Joshi, C. Kovar, M. Torralba, K. A. Wetterstrand, A. Abouellleil, A. M. Wollam, C. J. Buhay, Y. Ding, S. Dugan, M. G. FitzGerald, M. Holder, J. Hostetler, S. W. Clifton, E. Allen-Vercoe, A. M. Earl, C. N. Farmer, K. Liolios, M. G. Surette, Q. Xu, C. Pohl, K. Wilczek-Boney, and D. Zhu, A catalog of reference genomes from the human microbiome., Science, vol. 328, no. 5981, pp. 994999, May 2010.

[49] K. M. Wylie, R. M. Truty, T. J. Sharpton, K. A. Mihindukulasuriya, Y. Zhou, H. Gao, E. Sodergren, G. M. Weinstock, and K. S. Pollard, Novel bacterial

taxa in the human microbiome., PLoS ONE, vol. 7, no. 6, p. e35294, 2012.

[50] S. Chatterji, I. Yamazaki, Z. Bai, and J. A. Eisen, CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads, presented at the Research in Computational Molecular Biology, 2008, pp. 1728.

[51] J. L. Martinez, F. Baquero, and D. I. Andersson, Predicting antibiotic resistance, Nat. Rev. Microbiol., vol. 5, no. 12, pp. 958965, Dec. 2007.

[52] B. Bioinformatics, FastQC A quality control tool for high throughput sequence data, Cambridge, UK: Babraham Institute, 2011.

[53] B. Langmead and S. L. Salzberg, Fast gapped-read alignment with Bowtie 2., Nature Publishing Group, vol. 9, no. 4, pp. 357359, Apr. 2012.

[54] L. R. Meyer, A. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, and B. Rhead, The UCSC Genome Browser database: extensions and updates 2013, Nucleic Acids Res., vol. 41, no. 1, pp. D64D69, 2013.

[55] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing., J. Comput. Biol., vol. 19, no. 5, pp. 455477, May 2012.

[56] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool, Journal of Molecular Biology, vol. 215, no. 3, pp. 403410, 1990.

[57] A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models., Bioinformatics, vol. 22, no. 21, pp. 26882690, Nov. 2006.

[58] A. G. McArthur, N. Waglechner, F. Nizam, A. Yan, M. A. Azad, A. J. Baylay, K. Bhullar, M. J. Canova, G. De Pascale, L. Ejim, L. Kalan, A. M. King, K. Koteva, M. Morar, M. R. Mulvey, J. S. O'Brien, A. C. Pawlowski, L. J. V. Piddock, P. Spanogiannopoulos, A. D. Sutherland, I. Tang, P. L. Taylor, M. Thaker, W. Wang, M. Yan, T. Yu, and G. D. Wright, The Comprehensive Antibiotic Resistance Database, Antimicrobial Agents and Chemotherapy, vol. 57, no. 7, pp. 33483357, Jun. 2013.

[59] A. C. E. Darling, B. Mau, F. R. Blattner, and N. T. Perna, Mauve: multiple alignment of conserved genomic sequence with rearrangements., Genome Res., vol. 14, no. 7, pp. 13941403, Jul. 2004.

Chapter 5

Pathogens and host immunity in the ancient human oral cavity

As a part of this project, I was responsible for the analysis of whole metagenome shotgun sequence data collected from dental calculus of two samples G12 and B61 recovered from an excavation led by Dr. Warinner at an approximately 1000 year old burial site in Dalheim, Northern Germany. The aim of this study was to reconstruct health and dietary histories of the ancient individuals using a combination of 16S rRNA and WGS metagenomic, metaproteomic and microscopy based analysis. The list of my contributions is as follows.

1. Relative proportions of different bacteria, viruses, archea and fungi were estimated using the assembled whole metagenome sequencing data and a high confidence taxa list of all the identified microbes was prepared.
2. Using gene based evidence, 40 opportunistic pathogens were identified as inhabitants of the oral microbial community.
3. Virulence and drug resistance associated genes were identified to functionally characterize the metagenome of the ancient oral community.
4. Near complete whole genome reconstruction of a periodontitis pathogen *Tanerella forsythia* was carried out and subsequently analyzed for the presence of house keeping and virulence associated genes. Two pathogenicity associated islands were found to be missing that are present in the contemporary reference genome.
5. 100 whole metagenome shotgun datasets from contemporary plaque samples were analyzed to draw comparisons with the findings from ancient dental calculus.

Minor contributions include

1. Aided in construction of in house proteomic database and annotating of proteomic datasets using homology based approaches.

2. Assisted in identification of sequences with possible dietary origin using phylogenetic analysis

Summary

Dental Calculus, a calcified form of dental plaque is of great interest to archeologists since it offers a unique opportunity to examine the life style of ancient individuals. As plaque is accumulated over an entire life time in the absence of dental hygiene practices, it entraps debris from the oral and upper respiratory microbes (that would occasionally get coughed up in the oral cavity) and food particles. This information is sealed when the dental plaque calcifies post mortem to form a hard, impervious coat that preserves extremely well for over several centuries. In the current study, molecular sequencing techniques and microscopy has been used to investigate the impact of oral microbes on the health of ancient individuals. Several commensal and pathogenic microbes implicated in oral respiratory and cardiovascular diseases have been successfully identified. Genomic evidence indicating that the oral cavity has long been a reservoir for putative resistance genes was found and a near complete genome of a red complex periodontitis pathogen was reconstructed with missing genomic pathogenicity islands. Lastly, DNA sequences of plant and animal origin were identified providing clues about ancient dietary practices.

Pathogens and host immunity in the ancient human oral cavity

Christina Warinner[1,2], João F Matias Rodrigues[3,4], Rounak Vyas[3,4], Christian Trachsel[5], Natallia Shved[1], Jonas Grossmann[5], Anita Radini[6,7], Y Hancock[8], Raul Y Tito[2], Sarah Fiddyment[6], Camilla Speller[6], Jessica Hendy[6], Sophy Charlton[6], Hans Ulrich Luder[9], Domingo C Salazar-García[10–12], Elisabeth Eppler[13,14], Roger Seiler[1], Lars H Hansen[15,16], José Alfredo Samaniego Castruita[17], Simon Barkow-Oesterreicher[5], Kai Yik Teoh[6], Christian D Kelstrup[18], Jesper V Olsen[18], Paolo Nanni[5], Toshihisa Kawai[19,20], Eske Willerslev[17], Christian von Mering[3,4], Cecil M Lewis Jr[2], Matthew J Collins[6], M Thomas P Gilbert[17,21], Frank Rühli[1,22] & Enrico Cappellini[17,22]

Calcified dental plaque (dental calculus) preserves for millennia and entraps biomolecules from all domains of life and viruses. We report the first, to our knowledge, high-resolution taxonomic and protein functional characterization of the ancient oral microbiome and demonstrate that the oral cavity has long served as a reservoir for bacteria implicated in both local and systemic disease. We characterize (i) the ancient oral microbiome in a diseased state, (ii) 40 opportunistic pathogens, (iii) ancient human–associated putative antibiotic resistance genes, (iv) a genome reconstruction of the periodontal pathogen *Tannerella forsythia*, (v) 239 bacterial and 43 human proteins, allowing confirmation of a long-term association between host immune factors, 'red complex' pathogens and periodontal disease, and (vi) DNA sequences matching dietary sources. Directly datable and nearly ubiquitous, dental calculus permits the simultaneous investigation of pathogen activity, host immunity and diet, thereby extending direct investigation of common diseases into the human evolutionary past.

Unlike other human microbiomes, the oral microbiome will cause disease in a majority of people during their lifetime, suggesting that it is currently in a state of dysbiosis rather than symbiosis[1,2]. The human oral microbiome comprises more than 2,000 bacterial taxa, including a large number of opportunistic pathogens involved in periodontal, respiratory, cardiovascular and systemic diseases[3–7]. Dental calculus, a complex, calcified bacterial biofilm formed from dental plaque, saliva and gingival crevicular fluid[8], is emerging as a potential substrate for the direct investigation of the evolution of the oral microbiome and associated measures of oral health and diet[9,10]. Recently, a DNA-based 16S rRNA phylotyping study identified the major bacterial phyla in dental calculus and argued for shifts in microbial diversity associated with the origins of agriculture and industrialization[11], and, so far, five common oral bacteria have been identified in historic and prehistoric dental calculus using targeted PCR[12], quantitative PCR (qPCR)[11] and immunohistochemistry[13]. However, phylum-level community analysis and single-species targeted amplification are insufficient to characterize oral health and disease states, as this requires a deeper taxonomic and functional understanding of microbiome ecology[14].

We present the first detailed analysis to our knowledge of ancient oral microbiome ecology and function at the genus and species levels, leading to a deeper understanding of recent evolution of the human oral microbiome. Focusing on the dental tissues of four adult human skeletons (G12, B17, B61 and B78) with evidence of mild to severe periodontal disease from the medieval monastic site of Dalheim, Germany (c. 950–1200 CE) (**Supplementary Fig. 1**), as well as modern dental calculus from nine individuals with known dental histories, we demonstrate for the first time, to our knowledge, that the human oral microbiome has long served as a

[1]Centre for Evolutionary Medicine, Institute of Anatomy, University of Zürich, Zürich, Switzerland. [2]Department of Anthropology, University of Oklahoma, Norman, Oklahoma, USA. [3]Institute of Molecular Life Sciences, University of Zürich, Zürich, Switzerland. [4]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [5]Functional Genomics Center Zürich, University of Zürich/Swiss Federal Institute of Technology (ETH) Zürich, Zürich, Switzerland. [6]BioArCh, Department of Archaeology, University of York, York, UK. [7]University of Leicester Archaeological Services (ULAS), School of Archaeology and Ancient History, University of Leicester, Leicester, UK. [8]Department of Physics, University of York, York, UK. [9]Centre of Dental Medicine, Institute of Oral Biology, University of Zürich, Zürich, Switzerland. [10]Research Group on Plant Foods in Hominin Dietary Ecology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [11]Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [12]Department of Prehistory and Archaeology, University of Valencia, Valencia, Spain. [13]Research Group Neuro-Endocrine-Immune Interactions, Institute of Anatomy, University of Zürich, Zürich, Switzerland. [14]Zürich Center for Integrative Human Physiology, University of Zürich, Zürich, Switzerland. [15]Department of Biology, Microbiology, University of Copenhagen, Copenhagen, Denmark. [16]Department of Environmental Science, Aarhus Universitet, Roskilde, Denmark. [17]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. [18]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [19]Department of Immunology and Infectious Diseases, Forsyth Institute, Cambridge, Massachusetts, USA. [20]Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Harvard University, Boston, Massachusetts, USA. [21]Ancient DNA Laboratory, Murdoch University, Perth, Western Australia, Australia. [22]These authors jointly directed this work. Correspondence should be addressed to C.W. (twarinner@gmail.com) or E.C. (ecappellini@gmail.com).

Received 31 May 2013; accepted 3 February 2014; published online 23 February 2014; doi:10.1038/ng.2906

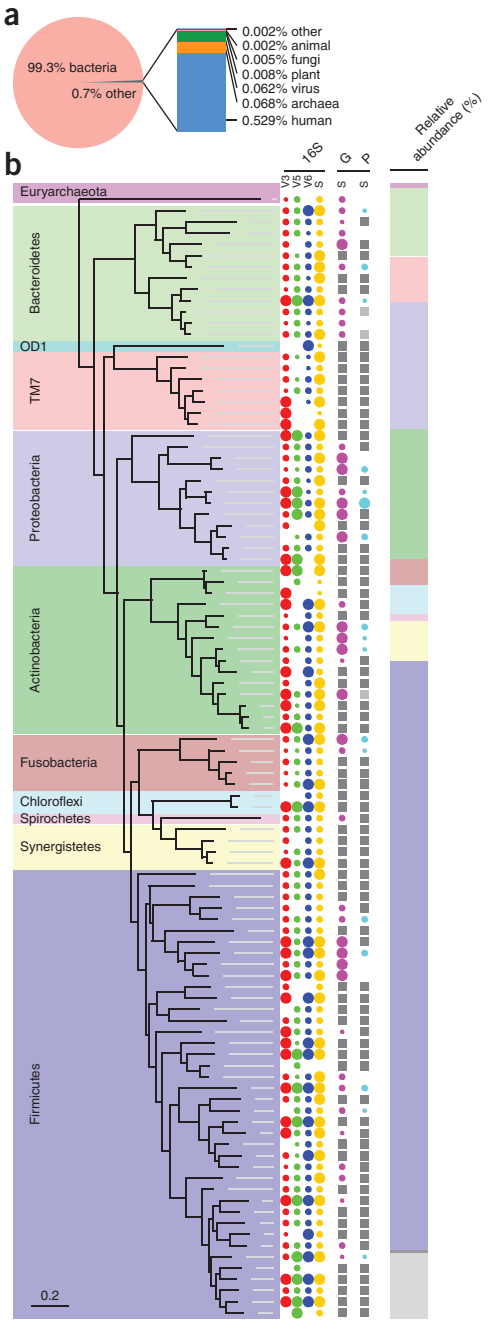

Figure 1 Taxonomic and phylogenetic characterization of ancient dental calculus. (**a**) Relative proportions of bacterial, archaeal, eukaryotic and viral DNA in ancient calculus estimated from assembled whole-metagenome shotgun sequences of two individuals. (**b**) Phylogenetic tree of the 100 most abundant OTUs in ancient dental calculus samples from 4 pooled individuals. Evidence for the presence and abundance of each microbial OTU is represented by colored, size-scaled circles for each targeted 16S rRNA hypervariable region (V3, V5, V6), shotgun 16S rRNA sequences (S) and other genes (G) and proteins (P) assigned to that OTU. OTUs for which no reference genome exists or for which insufficient proteome data have been validated for inclusion in the protein search databases are marked with gray squares, as no hits could be matched to those OTUs. The tree scale bar indicates nucleotide substitutions per site. Relative phylum abundance (normalized mean of all genetic data generated from 16S rRNA sequences) is represented by a column chart showing the phyla represented in the top 100 OTUs (colored), remaining phyla (dark gray) and unidentified OTUs (light gray).

reservoir for a broad suite of opportunistic pathogens implicated in both local and systemic disease and harbored a diverse range of putative antibiotic resistance genes. We confirm the long-term role of host immune activity and red complex pathogen virulence in periodontal pathogenesis, despite major changes in lifestyle, hygiene and diet over the past millennium. We reconstruct the genome of a major periodontal pathogen and present the first evidence, to our knowledge, of dietary biomolecules to be recovered from ancient dental calculus. Finally, we further validate our findings by applying multiple microscopic, genetic and proteomic analyses in parallel, providing a systematic biomolecular evaluation of ancient dental calculus preservation, taphonomy and contamination.

RESULTS

The ancient oral microbiome

Applying shotgun DNA sequencing to dental calculus for the first time, we found that it was strongly dominated by bacterial DNA, with minor contributions from human, viral, dietary and fungal sources (**Fig. 1a**). Using both targeted and shotgun 16S rRNA sequences (n = 509,067), we identified a total of 2,699 microbial operational taxonomic units (OTUs) in the ancient dental calculus, with the 100 most abundant taxa accounting for 86.6% of the total reads (**Fig. 1b** and **Supplementary Fig. 2**). One archaeal and nine bacterial phyla were dominant in ancient dental calculus (**Supplementary Table 1**): Firmicutes (49.5 ± 10.6%), Actinobacteria (12.0 ± 6.1%), Proteobacteria (11.5 ± 8.6%), Bacteroidetes (6.6 ± 3.6%), TM7 (4.6 ± 4.0%), Synergistetes (3.3 ± 2.6%), Chloroflexi (2.7 ± 1.5%), Fusobacteria (2.1 ± 1.8%), Spirochetes (0.6 ± 0.3%) and Euryarchaeota (0.4 ± 0.6%), all of which are also dominant in the human oral microbiome today[4]. Notably rare in ancient dental calculus was Acidobacteria, a ubiquitous and abundant bacterial phylum in soil[15].

To address biases resulting from the sequencing approach and the 16S rRNA gene hypervariable region (V3, V5, V6) primer choice (**Supplementary Fig. 3**), we visualized evidence for each OTU separately for each targeted and shotgun 16S rRNA detection method, as well as for shotgun metagenomic and metaproteomic data (**Fig. 1b**). Most OTUs were detected using multiple methods. OTUs detected from targeted V3 and shotgun data generally showed good agreement, whereas V5 and V6 primers showed clear evidence of primer bias and OTU dropout. Shotgun metagenomic data showed excellent agreement with consensus 16S rRNA OTUs when reference genomes were available. Shotgun metaproteomic data also showed good agreement with the OTUs identified on the basis of genetic data, and agreement is expected to improve as protein databases grow to include more predicted proteins and epigenetic variants. Because ancient DNA and proteins undergo different taphonomic processes and have different contamination risks, the high degree of phylogenetic consensus observed for data generated from independent extractions, using different methods and targeting different biomolecular types, demonstrates that an endogenous oral microbiome can be robustly and reliably recovered from ancient dental calculus.

Carriage of specific pathogens

The normal human oral flora includes a large number of endogenous cariogenic, periodontal and other opportunistic pathogens. Although these taxa generally do not cause extraoral disease in healthy subjects, they nevertheless pose a serious risk for the elderly and immunocompromised[16,17] and are known to be involved in the etiology of chronic systemic diseases, including cardiovascular disease[18]. As the detection of particular species from metagenomic sequence data is an open area of research, we applied a conservative contig assembly and BLAST strategy and screened our results against the Pathosystems Resource

Integration Center (PATRIC) database[19] to identify 40 putative opportunistic pathogens in ancient dental calculus (**Table 1**), of which only 5 had been previously reported in ancient samples[11–13]. We also identified phage DNA sequences specific to particular bacteria (**Table 1**), including *Streptococcus mitis* phage SM1, which has been previously shown to mediate *S. mitis* attachment to platelets and to increase bacterial virulence in the endocardium[20].

Both DNA and proteins from the periodontal pathogens *T. forsythia*, *Porphyromonas gingivalis* and *Treponema denticola* were particularly abundant in our ancient dental calculus samples, demonstrating that these so-called red complex bacteria[21] were strongly associated with periodontal disease during the medieval period, just as they are today, despite substantial changes in oral hygiene, diet and lifestyle. Additionally, all three of these pathogens were found at substantially higher frequency in our ancient dental calculus samples than in the Human Microbiome Project (HMP)[3] healthy cohort (**Supplementary Fig. 4a–c** and **Supplementary Table 2**), consistent with expectations for periodontal disease. We also identified several oral taxa (for example, *Aggregatibacter actinomycetemcomitans*, *Streptococcus mutans* and *S. mitis*) that have been shown to cause bacteremia and infective endocarditis[7,18], demonstrating that the human oral microbiome has long harbored pathogens that contribute to risk of cardiovascular disease. Additional pathogens included those implicated in acute dental infections (for example, *Actinomyces odontolyticus*), caries (*S. mutans*) and opportunistic upper and lower respiratory illness (for example, *Streptococcus pneumoniae*, *Streptococcus pyogenes* and *Haemophilus influenzae*). Of interest, all ancient dental calculus samples were also found to contain disordered carbon (microcharcoal), a respiratory irritant. Two obligate human taxa, *Neisseria meningitidis* and *Neisseria gonorrhoeae*, causative agents of bacterial meningitis and gonorrhea, respectively, were also observed. *N. meningitidis* and *N. gonorrhoeae* form a recently diverged pathogenic clade of *Neisseria*, a genus comprising many commensal species inhabiting the mucosa and dental surfaces of animals[22], and both are prevalent members of the human oral microbiome. Genital *N. gonorrhoeae* strains can infect the pharynx and engage in genetic exchange with other *Neisseria* species;[23] however, oral strains are not known to cause genital infection. Oral *N. meningitidis* is a leading cause of bacterial meningitis, although disease susceptibility is determined by a combination of host genetics and strain virulence[24]. Finally, we observed two additional oral taxa present at substantially higher frequency in at least one ancient dental calculus sample compared to the HMP healthy cohort: *Filifactor alocis* and *Olsenella uli* (**Supplementary Fig. 4e,f**). Although not classified as pathogens in the PATRIC database, these bacteria have recently been associated with periodontitis and endodontic infections, respectively[25,26].

Table 1 Putative pathogens identified from assembled metagenomic and metaproteomic sequences in ancient dental calculus

Pathogens[a]	Genes (contigs)	Proteins (peptides)	Virulence	Drug resistance[b]	Plasmid	CTn or phage
Actinomyces odontolyticus[c]	3 (4)	3 (34)				
Aggregatibacter actinomycetemcomitans	50 (68)	0			+	+
Campylobacter concisus	10 (20)	0			+	
Campylobacter curvus	12 (11)	0				
Campylobacter rectus[c]	3 (9)	3 (15)	++			
Campylobacter showae[c]	3 (13)	1 (2)				
Capnocytophaga gingivalis[c]	2 (11)	3 (7)	+			
Capnocytophaga ochracea	938 (4,909)	0	+	+	+	+
Capnocytophaga sputigena[c]	2 (2)	0		+		
Clostridium difficile[d,e]	30 (76)	0		+		+
Corynebacterium matruchotii[c]	2 (15)	12 (89)				
Eikenella corrodens[c]	11 (38)	2 (11)			+	
Fusobacterium nucleatum	656 (1,525)	4 (21)	++	+	+	+
Fusobacterium periodonticum[c]	3 (6)	3 (8)	+			
Gemella morbillorum[c]	9 (38)	0				
Gordonibacter pamelae[d]	3 (30)	0				
Haemophilus influenzae	19 (43)	1 (4)				+
Histophilus somni[d,f]	9 (18)	0				+
Leptotrichia buccalis	492 (1,104)	0	+	+	+	+
Neisseria gonorrhoeae	127 (250)	1 (2)	+		+	
Neisseria meningitidis	336 (821)	1 (2)	+	+		+
Neisseria sicca[c]	3 (8)	4 (35)				
Neisseria subflava[c]	4 (12)	0				
Porphyromonas gingivalis	802 (2,588)	7 (72)	++	+		+
Rothia mucilaginosa	24 (17)	1 (2)		+		
Streptobacillus moniliformis[d,f]	8 (23)	0				
Streptococcus agalactiae	7 (27)	0			+	+
Streptococcus dysgalactiae[d]	2 (8)	0				+
Streptococcus equi[d,f]	29 (101)	0	+			+
Streptococcus gallolyticus[d,f]	8 (11)	0				+
Streptococcus gordonii	882 (3,397)	1 (8)	+	+	+	+
Streptococcus mitis	88 (161)	1 (37)	++	+		+
Streptococcus mutans	21 (67)	0				
Streptococcus pneumoniae	144 (339)	1 (8)	+	+		+
Streptococcus pyogenes	14 (32)	1 (8)		+		+
Streptococcus sanguinis	850 (3,272)	1 (4)	+	+		
Streptococcus suis[d,f]	2 (3)	0	+			
Tannerella forsythia	1,099 (11,279)	10 (137)	++	+		+
Treponema denticola	917 (6,106)	3 (15)	++	+	+	+
Veillonella parvula	96 (109)	0		+		

Metagenomic data from G12 and B61 and proteomic data from G12, B17, B61 and B78. +, gene(s) detected; ++, gene(s) and protein(s) detected.

[a]Includes only pathogens with valid entries in the PATRIC database. Only taxa represented by more than one DNA contig are shown. All pathogens are known inhabitants of the human oral cavity, as confirmed by cross-referencing with HMP data for supragingival dental plaque. [b]Putative function based on gene homology and NCBI annotation; functionality was not independently validated. [c]Reference genome sequencing and annotation incomplete. [d]Not a prevalent inhabitant of the oral cavity; not present in the Human Oral Microbiome Database (HOMD). [e]Tentative identification; sequences correspond almost exclusively to mobile genetic elements. [f]Putative zoonosis.

Virulence

To further characterize the pathogens detected in ancient human dental calculus, we compared information on the functional features of putative genes and proteins associated with virulence, drug resistance, plasmids, transposons and phages to the information available in NCBI records. Although not exhaustive, a preliminary list of well-supported virulence-associated genes and proteins was compiled

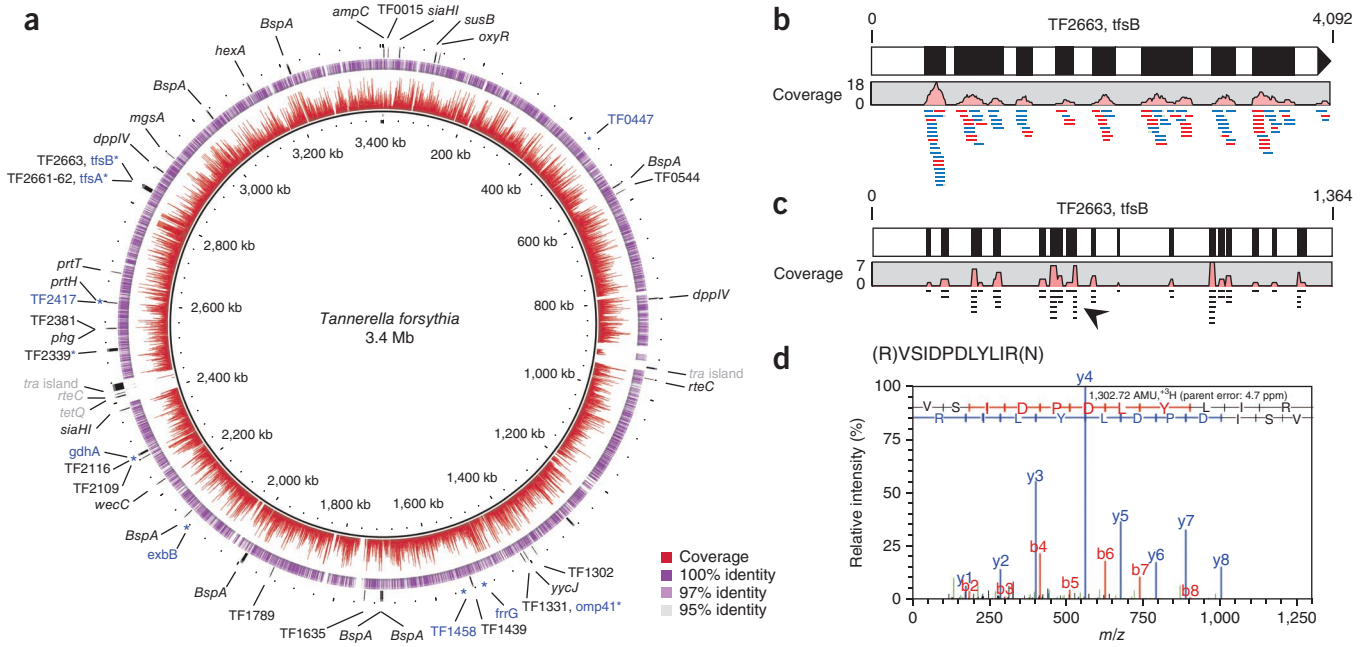

Figure 2 Genomic coverage plot for the periodontal pathogen *T. forsythia*, with details of gene and protein coverage of the virulence factor TF2663, tfsB, from medieval human dental calculus (G12). (**a**) Plot of the *T. forsythia* genome with depth of coverage (0- to 30-fold) shown in red and identity shown in purple. Gene locations of the major virulence factors present (black) or absent (gray) in the assembly are indicated in the outer ring. Notably absent are two transposon-related *tra* pathogenicity islands containing putative tetracycline resistance genes in the *T. forsythia* ATCC 43037 reference strain. Proteins identified by MS/MS are indicated in blue by an asterisk at the corresponding gene locus. (**b**) Enlarged view of forward (blue) and reverse (red) DNA reads mapped to gene TF2663, which encodes the surface layer B protein tfsB. (**c**) Enlarged view of peptide coverage for tfsB, a species-diagnostic virulence factor involved in hemagglutination, adherence and host tissue invasion. (**d**) Example of a tandem mass spectrum indicating the measured mass-to-charge ratios (*m*/*z*) of the fragments supporting the identification of the tfsB peptide VSIDPDLYLIR (arrowhead in **c**).

using this method (**Table 1** and **Supplementary Table 3**), showing a wide range of virulence factors associated with adhesion and aggregation (for example, adhesins and lectins) and parasitism (for example, phospholipases, hemagglutinins and hemolysins), as well as extensive machinery for horizontal gene transfer (for example, pilin, CTn and phage sequences). In several cases, we detected both the virulence-associated gene and its protein product, for example, *Msp* and major sheath protein in *T. denticola* and *Rgp* and Arg-gingipain in *P. gingivalis*. Arg-gingipain and Lys-gingipain, another extracellular cysteine proteinase identified by proteomic evidence, are highly antigenic and extremely abundant in *P. gingivalis*, accounting for 10% of the total protein produced by the organism[27]. Notably, we also detected type IV fimbrilin, an outer membrane protein variant associated with virulent *P. gingivalis* strains[28].

Antibiotic resistance

The human microbiome is an important site of horizontal gene transfer and a potential reservoir of antimicrobial resistance[29]. Metagenomic studies of modern dental plaque have found a wide range of predicted genes related to resistance to diverse antibiotics and toxic compounds[30]. The antiquity of bacterial antibiotic resistance genes has recently been tested in permafrost soils dating to the Pleistocene[31], but, until now, the antiquity of antibiotic resistance in human microbiota before the use of therapeutic antibiotics had not been investigated.

Using both automated and manual search strategies, we identified within ancient dental calculus numerous DNA sequences with homology to antibiotic resistance genes found in oral and pathogenic bacteria, including genes for multidrug efflux pumps and native resistance genes to aminoglycosides, β-lactams, bacitracin, bacteriocins and macrolides, among others, as well as a near-complete plasmid-encoded conjugative transposon carrying efflux pump genes with high homology to CTn5 of *Clostridium difficile* (**Supplementary Table 4**). Although the exact function of these genes in our samples is unclear, their presence nevertheless demonstrates that the biomolecular machinery for broad-spectrum, low-level antibiotic resistance has long been present in the human microbiome, illustrating how the oral microbiome functions as both a source and a reservoir of new antibiotic resistance[23].

Pathogen genome reconstruction: *T. forsythia*

T. forsythia (formerly *Bacteroides forsythus* and *Tannerella forsythensis*) is an anaerobic, Gram-negative member of the phylum Bacteroidetes and is a known inhabitant of supragingival and subgingival plaque[32]. It is associated with advanced forms of periodontal disease and has been reported in atherosclerotic lesions[7]. On the basis of 16S rRNA gene data, *T. forsythia* was observed to be at moderate abundance (0.09–0.84%) in the dental calculus of one individual (G12) and, as a pathogen of interest, was selected for genome reconstruction.

Using a conservative mapping strategy, a total of 10,991 contigs were recruited to the ancient *T. forsythia* genome reconstruction, at a mean nucleotide depth of coverage of 5.7 (**Fig. 2a**). Ninety-one percent of *T. forsythia* genes ($n = 2{,}799$) were mapped by at least one contig, and unmapped genes included 94 transposases, transfer factors and other mobilization genes that may be specific to the *T. forsythia* ATCC 43037 reference genome strain used for alignment. The largest gap in our genome reconstruction, which spanned ~48,000 bp and 53 genes, corresponded to a complete conjugative transposon carrying putative tetracycline resistance genes that was absent in our reconstructed ancient *T. forsythia* genome. In addition to genetic sequences, tandem

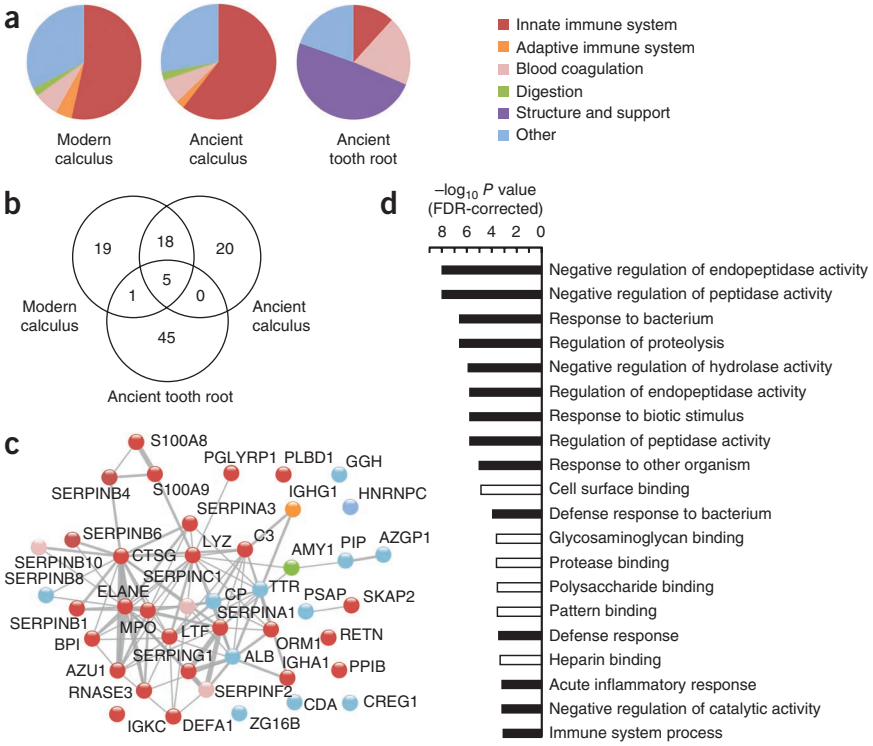

Figure 3 Metaproteomic comparison of human proteins in modern and ancient dental samples. (**a**) Functional characterization of human proteins identified in modern dental calculus (two individuals), ancient dental calculus (four individuals) and ancient tooth roots (four individuals). (**b**) Venn diagram of shared human proteins by sample type. (**c**) STRING network representation of human proteins identified in ancient dental calculus; nodes are labeled by protein name and colored in accordance with functional categories and connections (gray) to predicted functional partners. The network is set to medium confidence (0.4) for all active prediction methods. (**d**) Gene ontology categories with significant enrichment ($P < 0.001$, FDR-corrected) in ancient dental calculus for biological processes (black) and molecular functions (white) related to proteinase regulation, substrate binding and innate immune function. Enrichment is calculated relative to the human genome using the STRING-embedded AmiGO term enrichment tool.

mass spectrometry (MS/MS) identified 118 peptides belonging to 10 *T. forsythia* proteins (**Fig. 2a**). Of these proteins, nine were outer membrane or S-layer proteins, seven had a known function and four were antigenic: *T. forsythia* surface protein A (TF2661-2, tfsA), *T. forsythia* surface protein B (TF2663, tfsB), outer membrane protein 41 (TF1331, omp41) and one hypothetical protein (TF2339)[33].

Several virulence factors and antigenic proteins have been identified in *T. forsythia* (**Fig. 2a**), including Bacteroides surface protein A (BspA), dipeptidyl peptidase-4 (dppIV), tfsA and tfsB, among others[27,33]. The genes encoding each of these virulence factors were present in our reconstruction. The glycosylated *T. forsythia* S-layer proteins tfsA and tfsB are directly involved in hemagglutination, adhesion and tissue invasion[34]. They are also unique and are species diagnostic, as they have no homology to other known S-layer proteins or glycoproteins[35]. DNA and protein coverage of tfsA and tfsB was high in our data set; for example, 10 contigs comprising 116 reads mapped to the TF2663 (*tfsB*) gene (**Fig. 2b**), and we identified 65 spectra belonging to 27 unique tfsB peptides (**Fig. 2c,d**). Given that functional *T. forsythia* S-layer protein is essential for host immune evasion and biofilm coaggregation[34], the discovery of abundant, well-preserved S-layer gene and protein sequences makes ancient dental calculus an excellent candidate for investigating the evolution of periodontal pathogenesis in humans.

MS/MS analysis of host immunity and disease pathogenesis

Despite dense microbial colonization and the regular introduction of foreign substances, the oral cavity is effective at preventing most infections. At least 45 antimicrobial gene products acting as early responders of the innate immune system have been identified in saliva and gingival crevicular fluid[36]. We identified 43 human proteins within ancient dental calculus, of which 25 are involved in the innate immune system (**Fig. 3a**). Eight of these proteins have demonstrated antimicrobial properties and include cationic peptides (α-defensin and azurocidin), metal ion chelators (calgranulin A, calgranulin B and lactoferrin), protease inhibitors (myeloperoxidase) and bactericidal proteins (bactericidal permeability-increasing protein, lysozyme C and peptidoglycan recognition protein 1). Expression of many of these proteins is specific to a particular cell type and even a particular subcellular component (for example, azurocidin is specific to neutrophil lysozomic azurophilic granules), allowing highly resolved characterization of the immune system response. Approximately one-third of the identified human proteins were shared by ancient and modern calculus (**Fig. 3b**), and functional profiles were highly similar (**Fig. 3a**). In contrast, ancient tooth roots were distinct, both in protein composition and function and being dominated by collagens and other proteins involved in mineralized tissue (biglycan, periostin) and vascular (prothrombin) development and maintenance.

The STRING resource[37] was used to investigate functional interaction networks among the human proteins in ancient dental calculus. A large number of functional interactions were predicted (**Fig. 3c**), and 79% of proteins ($n = 34$) were functionally connected to at least one other protein in the network. Immunoglobulin heavy chain (IgA and IgG) and light chain (kappa) peptides were detected in ancient calculus, as was α-amylase, a salivary enzyme that breaks down dietary starch; however, the majority of proteins were related to the innate immune system. Human proteins in ancient dental calculus were strongly enriched in extracellular (P value of 3.2×10^{-12}, false discovery rate (FDR)-corrected) and secretory (P value of 4.3×10^{-9}, FDR-corrected) proteins, mostly of neutrophilic origin. Extravasated neutrophils are recruited to sites of injury by IgG and have a life span of less than 24 h[38]; thus, neutrophil proteins are only released into calcifying dental plaques during active infection and inflammation. Relatively few human cellular proteins were found, suggesting that immune cells do not invade the calcifying plaque but rather release antimicrobial substances from the junctional and pocket epithelia, a process that is consistent with neutrophil 'frustrated phagocytosis' (ref. 39) and NETosis[40]. Human proteins in ancient calculus were significantly enriched in biological processes related to inflammation, innate immunity and host defense, as well as in molecular functions such as cell surface, protease and glycosaminoglycan binding (**Fig. 3d**). The observation of an abundance of inflammatory (myeloperoxidase, azurocidin, lysozyme, calprotectin and elastase) and anti-inflammatory (α-1-antitrypsin and α-1-antichymotrypsin) innate immune system

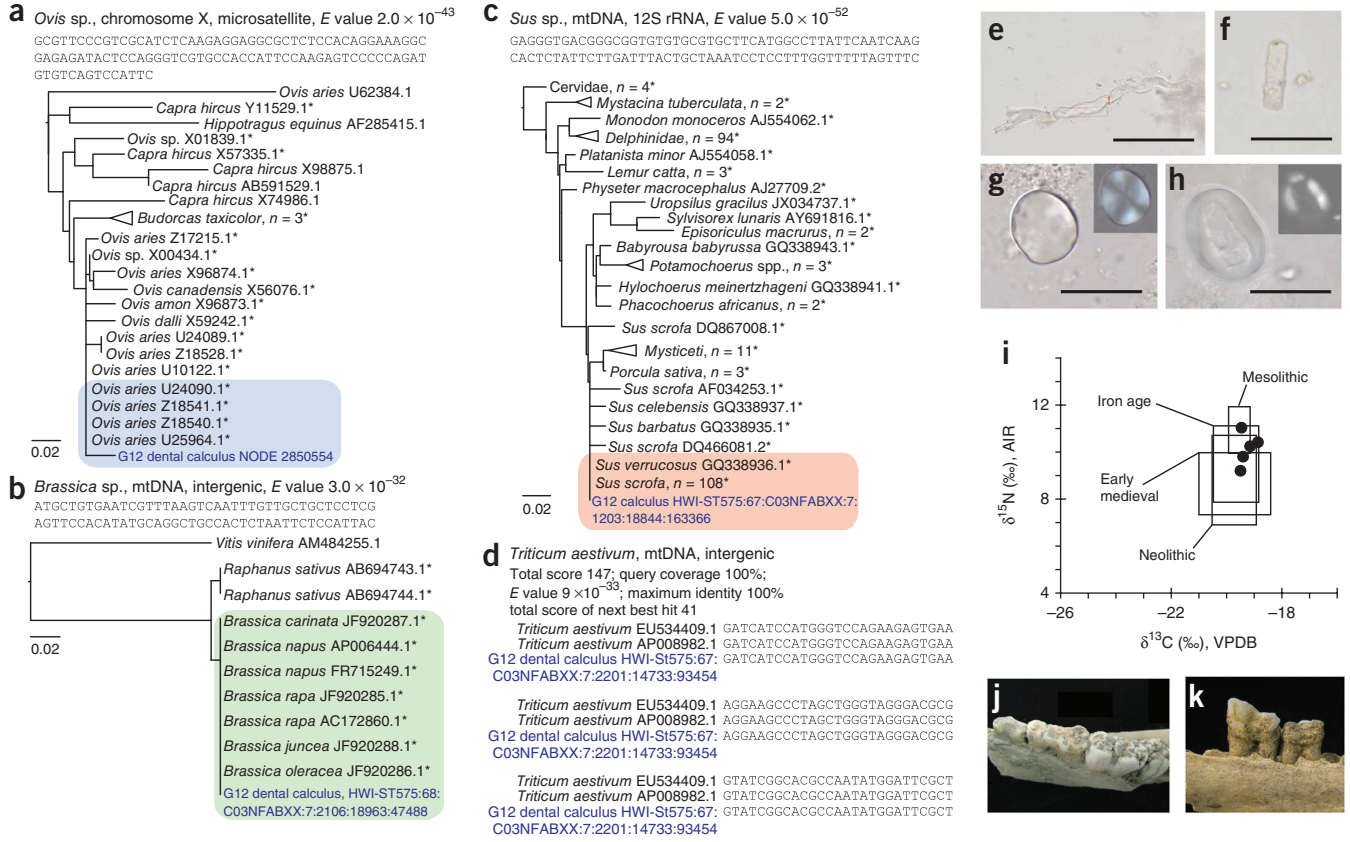

Figure 4 Genetic, microfossil, zooarchaeological and stable isotopic evidence for the medieval human diet at Dalheim, Germany. (**a**–**c**) Neighbor-joining trees for GenBank sequences aligning to putative dietary sheep (**a**), crucifer (**b**) and pig or boar (**c**) sequences. Trees include accessions with alignment scores of >45, except for **c**, which is limited to the top 250 alignments; highly significant alignments (E value < 1×10^{-30}) are indicated with an asterisk. BLAST top hits for each dietary sequence are highlighted. The maximum fraction of mismatched bases is 0.75 for tree generation, and distance was calculated using a Jukes-Cantor substitution model. Tree scale bars indicate nucleotide substitutions per site. mtDNA, mitochondrial DNA. (**d**) One sequence aligned to two accessions of bread wheat only. (**e**–**h**) The microfossils recovered from ancient human dental calculus yielded morphological matches to animal collagen fibers (**e**), a smooth long-cell phytolith (**f**) and starch granules of the grass tribe Triticeae (**g**) and the legume family Fabaceae (**h**). Characteristic starch granule birefringence is shown under polarized light in the insets in **g**,**h**. Scale bars, 50 µm in **e**, 25 µm in **f**–**h**. (**i**) C and N stable isotopic values for human bone collagen (black circles) fall within 2 s.d. (boxes) of those measured for other Central European populations and are consistent with a diet of mixed C_3 terrestrial plant and animal resources. Isotopic values are reported in delta notation: $\delta^{15}N = (^{15}N/^{14}N)_{sample}/(^{15}N/^{14}N)_{AIR} - 1$, with the air (AIR) standard; $\delta^{13}C = (^{13}C/^{12}C)_{sample}/(^{13}C/^{12}C)_{VPDB} - 1$, with the Vienna Pee Dee Belemnite (VPDB) standard. (**j**,**k**) Recovered food waste includes skeletal material from *Sus* species (**j**) and Caprinae (**k**).

proteins in ancient dental calculus, coupled with morphological evidence of attachment loss and alveolar recession, is strongly supportive of active periodontal inflammation and disease.

In addition to this host immunological data, we identified oral pathogens and bacterial virulence proteins in ancient and modern dental calculus known to provoke strong immunological reaction and to contribute to periodontal pathogenesis (**Supplementary Table 3**), most notably *P. gingivalis* (gingipains), *T. forsythia* (S-layer proteins) and *T. denticola* (major sheath protein). *P. gingivalis* has recently been shown to stimulate neutrophils to release resistin, a protein implicated in acquired insulin resistance[41]. Resistin may exacerbate the progression of type 2 diabetes[42], and, interestingly, we identified resistin on the basis of reasonably abundant evidence (36 spectra, 9 unique peptides) in ancient dental calculus. Resistin was also identified in modern calculus (seven spectra, five unique peptides) but not in ancient tooth roots.

Ancient dietary reconstruction

Given current challenges in nutritional health and obesity[43], a growing interest in dietary aspects of the hygiene hypothesis[44] and a recent study suggesting shifts in the ancient oral microbiome associated with periods of agricultural transition[11], there is great interest in better understanding the evolutionary history of the human diet. However, paleodietary reconstruction is made difficult by the generally poor preservation of plants and small animals in the archaeological record. Stable isotope analysis of human bone and dental calculus–based plant microfossil research have broadened our knowledge of past dietary practices, but these tools are insufficient to characterize many major dietary components at high taxonomic resolution. Ancient DNA–based approaches offer great advantages and have been used to identify dietary components from archaeological feces (coprolites), as well as to investigate plant remains directly[45]. However, as coprolites and preserved plant remains are relatively rare, we sought to characterize dietary information from dental calculus using both biomolecular and conventional methods.

From our metagenomic sequence reads, a total of 487 reads (0.0003%) were confidently identified as eukaryotic organelle sequences; of these, 266 were assigned to the kingdom Viridiplantae, and 21 were assigned to the kingdom Animalia. Within these kingdoms, most of the organelle reads mapped ambiguously to multiple organisms or genera, leaving only 20 reads that could be positively identified at a subfamily level. Of these 20 reads, 17 were of host origin, and the remaining 3 reads

matched diagnostic mitochondrial sequences for pig/boar (*Sus* species), crucifer (*Brassica* species) and bread wheat (*Triticum aestivum*). Analysis of assembled contigs additionally identified one putative sheep (*Ovis* species) and several human ($n = 326$) nuclear genomic sequences (**Fig. 4a–d**). Although previous studies have reported trace animal domesticate DNA contamination (from cattle, pig and chicken) in some PCR reagents[46], we found no evidence of such contamination, and additionally wheat, crucifers and sheep are not part of this supply chain. The discovery of preserved dietary biomolecules is consistent with previous observations of intact dietary microfossils, such as starch granules, in archaeological dental calculus[10] and with reports of wheat and cassava (tapioca) chloroplast DNA in the dental plaque of living subjects[4]. Turning to proteins, we identified one putative dietary plant protein, chloroplast glyceraldehyde 3-phosphate dehydrogenase (GAPDH), in ancient calculus, but disambiguation below the phylum Viridiplantae was not possible. Faunal proteins were not confidently identified within ancient dental calculus, but we did identify bovine β-lactoglobulin, a milk protein, in modern dental calculus, demonstrating that recovery of dietary animal proteins from dental calculus is possible.

Because our discovery of dietary biomolecules in dental calculus is new, we sought to validate our results using independent paleodietary methods. Microfossil analysis of ancient dental calculus yielded morphological matches to animal connective tissue fragments ($n = 2$; **Fig. 4e**), an unidentified monocot phytolith (**Fig. 4f**), plant bast fibers ($n = 3$) and starch granules consistent with the cereal tribe Triticeae ($n = 27$; **Fig. 4g**) and the legume family Fabaceae ($n = 1$; **Fig. 4h**), among other debris (**Supplementary Fig. 5**). Stable isotope analysis of human bone collagen (**Fig. 4i**) from the four ancient human individuals indicated a mixed diet of C_3 terrestrial plant and animal resources typical of Central European populations from the late Mesolithic through the medieval period[47–50]. Zooarchaeological analysis of food waste at the site confirmed the presence of pig or boar (*Sus* species; **Fig. 4j**) and sheep or goat (Caprinae; **Fig. 4k**), as well as cattle (*Bos* species) and equids (*Equus* species).

Biomolecular analysis of dental calculus thus yields complementary dietary information compared to conventional methods, as well as new findings. The high taxonomic precision of genetic approaches allows closely related taxa (for example, Caprinae) to be distinguished in the absence of diagnostic skeletal elements, and under-represented plant taxa, such as *Brassica*, can be identified without the biological and taphonomic biases that compromise macro- and microfossil preservation of leafy greens and vegetables.

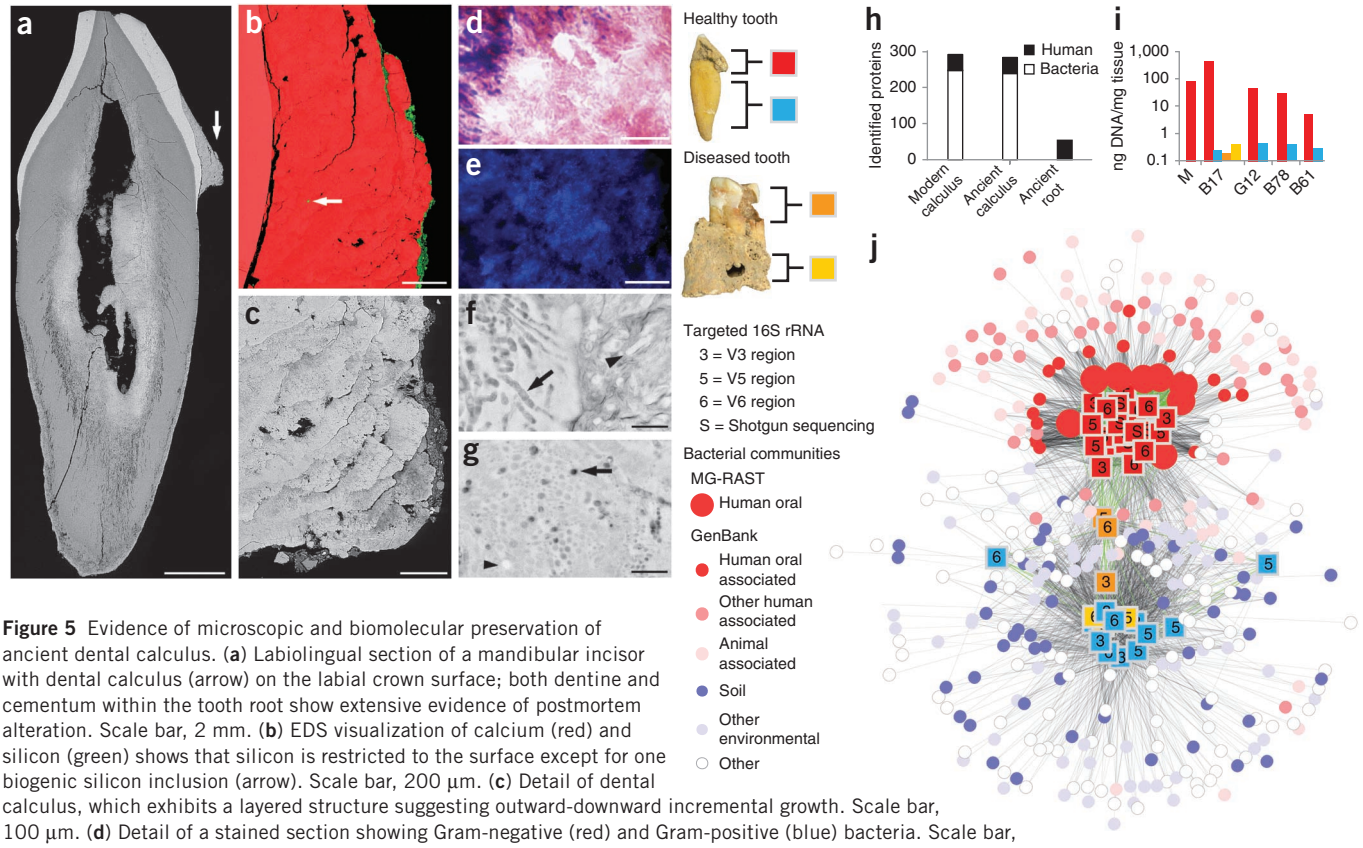

Figure 5 Evidence of microscopic and biomolecular preservation of ancient dental calculus. (**a**) Labiolingual section of a mandibular incisor with dental calculus (arrow) on the labial crown surface; both dentine and cementum within the tooth root show extensive evidence of postmortem alteration. Scale bar, 2 mm. (**b**) EDS visualization of calcium (red) and silicon (green) shows that silicon is restricted to the surface except for one biogenic silicon inclusion (arrow). Scale bar, 200 µm. (**c**) Detail of dental calculus, which exhibits a layered structure suggesting outward-downward incremental growth. Scale bar, 100 µm. (**d**) Detail of a stained section showing Gram-negative (red) and Gram-positive (blue) bacteria. Scale bar, 10 µm. (**e**) Detail of a Hoechst-stained section showing abundant *in situ* double-stranded DNA. Scale bar, 20 µm. (**f**) Detail of the calculus matrix containing numerous lacunae of filamentous microorganisms (arrow), some of which are mineralized (arrowhead). (**g**) In another area, the matrix contains both unmineralized (arrow) and mineralized (arrowhead) lacunae of microorganisms resembling cocci. Scale bars, 5 µm. (**h**) Proportions of human and bacterial proteins identified in ancient and modern samples. (**i**) DNA extraction yields from modern (M) and ancient dental calculus (red), dentine (blue and orange) and alveolar bone samples (yellow). (**j**) Comparison of microbial communities in ancient dental samples (squares) to those in diverse publicly available samples (circles). 16S rRNA data were generated by shotgun sequencing (S) and targeted amplification and sequencing of hypervariable regions (V3, V5, V6), followed by OTU clustering. Ancient metagenomes ($n = 38$) were plotted within a network where distance scales with OTU community similarity. Modern metagenomes with >20% shared community are shown connected by black lines, and ancient metagenomes with >20% shared community are connected by green lines. A total of 315 GenBank studies were recruited to the network. Ancient metagenomes segregate into two distinct groups: ancient dental calculus samples cluster tightly together, are connected by thick lines and show similarity to modern metagenomes of primarily human and oral origin, and ancient dentine and abscessed bone tissue samples form a more diffuse cluster and recruit primarily soil and environmental metagenomes. Carious dentine forms an intermediate cluster that shares OTUs with both human-associated and environmental sources.

Taphonomy and contamination

Postmortem taphonomy and contamination pose challenges in ancient biomolecular research. To address these potential problems in our data set, we employed multiple protocols for authenticating our data, including scanning electron microscopy (SEM), energy-dispersive X-ray spectroscopy (EDS), optical microscopy, Raman spectroscopy, protein damage analysis, genetic network analysis and probabilistic genetic source tracking.

After death, environmental microbes are known to infiltrate the dentition, causing substantial tissue degradation, loss of organic matter and altered mineralization patterns in dentine and cementum (**Supplementary Fig. 6**)[51]. We observed, however, little evidence of postmortem alteration in ancient dental calculus samples (**Fig. 5a**). EDS imaging showed a thin deposit of silicon-rich soil matrix only on the dental calculus surface (**Fig. 5b**), and no evidence of altered mineralization was observed within ancient dental calculus, a finding that we confirmed by Raman spectroscopic comparison with modern controls (**Supplementary Fig. 7**). During life, growth of dental calculus is appositional[8,52], resulting in a laminar cross-sectional structure characterized by alternating bands of Gram-positive and Gram-negative bacteria (**Supplementary Fig. 8**), a pattern we also observed in ancient calculus (**Fig. 5c**,**d**). DNA fluorescent dye showed a similar distribution of double-stranded DNA in ancient and modern calculus, in many cases resolving to individual cells (**Fig. 5e** and **Supplementary Fig. 9**) corresponding to a diverse range of *in situ* bacteria embedded within undisturbed dental calculus matrix (**Fig. 5f**,**g**).

Ancient dental calculus yielded microbial (n = 239) and human (n = 43) proteins in the same relative proportion and with similar functions as in modern controls (**Figs. 3a** and **5h**, and **Supplementary Fig. 10**), whereas only human proteins (n = 53) were confidently identified from tooth roots and bone. Damage analysis of dental calculus proteins showed a higher proportion of spontaneous, non-enzymatic post-translational modifications in ancient samples compared to modern controls; however, both modern and ancient dental calculus peptides exhibited relatively high proportions of non-tryptic cleavage (>10%), an observation consistent with *in vivo* exposure to bacterial and immune system proteases (**Supplementary Fig. 11**).

Total DNA recovery from ancient dental calculus (5–437 ng DNA/mg calculus) was comparable to that for modern calculus and one to three orders of magnitude greater than from paired dentine (0.3–0.5 ng/mg), carious dentine (0.2 ng/mg) and abscessed bone (0.4 ng/mg) (**Fig. 5i**). Analysis of 16S rRNA phylotypes using a new network analysis tool developed for this study showed that the bacterial communities within ancient dental calculus closely resembled published human oral microbiomes and were distinct from the communities observed in ancient dentine and bone, which clustered primarily with published soil samples, indicating environmental contamination after death (**Fig. 5j** and **Supplementary Fig. 12**). This pattern was found to be robust to extraction method, decontamination method, primer selection, sequencing method and interindividual variation. Reanalysis of our data using the methods employed by HMP[3] yielded equivalent results (**Supplementary Figs. 13–15**) that were also confirmed using the Bayesian tool SourceTracker[53] (**Supplementary Fig. 16**). Ancient dental calculus is thus shown to be a remarkably well-preserved biological material that allows direct and detailed investigations of the ancient oral microbiome.

DISCUSSION

Dental calculus is among the richest biomolecular sources yet identified in the archaeological record. Given the exceptional preservation of DNA within dental calculus (5–437 ng/mg), next-generation shotgun sequencing libraries can be built from milligrams of material, thereby reducing typical sample requirements for ancient DNA analysis by two orders of magnitude. We demonstrate that the red complex pathogens *T. forsythia*, *P. gingivalis* and *T. denticola* have long been associated with periodontal disease, despite changes in lifestyle, hygiene and diet since the medieval period. We confirm the long-term carriage of opportunistic pathogens in the human oral cavity, including the causative agents of oral and respiratory diseases, as well as bacteria implicated in the progression of cardiovascular disease and the formation of arterial plaques. We find genetic evidence that the human oral cavity has long harbored genes with homology to putative antibiotic resistance genes, the first such demonstration, to our knowledge, in an ancient human–associated sample. We reconstruct the genome of the periodontal pathogen *T. forsythia* without previous enrichment and identify the absence of a complete conjugative transposon carrying putative tetracycline resistance genes found in the reference strain. We report for the first time, to our knowledge, the presence of well-preserved proteins within ancient dental calculus and show that, although the dental calculus metagenome is dominated by bacterial DNA (>99%), the dental calculus metaproteome contains high proportions of both host and microbial proteins of clinical significance. Because the growth of calculus is appositional without remodeling, it may offer a potential solution to the 'osteological paradox' in studies of ancient disease[54], and, given that proteins are known to survive longer in the archaeological record than DNA, dental calculus may allow the recovery of valuable proteomic data from deep time periods that are out of reach using genomic technologies. Finally, we report the first plant and animal DNA sequences recovered from ancient dental calculus; these sequences allow greater taxonomic precision than is currently possible using microfossil or stable isotope paleodietary techniques. Dental calculus is a robust, long-term biomolecular reservoir of ancient disease and dietary information, and it has important implications for the fields of medicine, microbiome research, archaeology and human evolutionary studies.

URLs. Graphviz package, http://www.graphviz.org/; GitHub, https://github.com/jfmrod/metagenome-sample-network-generator.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Illumina and 454 genetic data have been deposited in the NCBI Short Read Archive (SRA) under the project accession SRP029257 and sample accessions SRS473742–SRS473771 and SRS480529–SRS480539 and to MG-RAST[55] under project accession 365 and accessions 4486524.3, 4486533.3, 4486537.3, 4486539.3, 4486540.3, 4486544.3, 4486613.3, 4486614.3, 4486617.3, 4487224.3–4487231.3, 4487233.3–4487235.3, 4487237.3–4487248.3, 4488534.3–4488536.3, 4488542.3, 4517539.3, 4530391.3, 4530438.3, 4530439.3 and 4530473.3–4530475.3. Proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository[56] with the data set identifier PXD000412 and accessions 34605–34628. Computer source code for the network analysis in **Figure 5j** has been deposited to GitHub (see URLs).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the Kantonale Ethik-Kommission Zürich, the Functional Genomics Center Zürich, the Center for Microscopy and Image Analysis, and the Institute of Oral Biology at the University of Zürich; the PRIDE Team; G. Akgül, K. Alt,

D. Ashford, P. Ashton, H. Barton, A. Bouwman, C. Burger, D. Coulthard, J. Hublin, V. Meskenaite, F. Najar, M. Richards, K. Sankaranarayanan, R. Schlapbach, L. Shillito, T. Stöllner, O. Ullrich and H. Zbinden for assistance with data collection, analysis and management; and M. Carver, F. Dewhirst, A. Tanner, K. Hardy and A. Henry for helpful comments on early drafts and data analyses. This work was supported by the Mäxi Foundation Zürich, the Swiss Foundation for Nutritional Research, Danish Research Foundation grant 29396, Danish Council for Independent Research grant 10-081390, Lundbeck Foundation grants R52-A5062 and R44-A4384, US National Institutes of Health grants R01-HG005172, R01-GM089886, R01-DE018499 and R21-DE018310, European Research Council grant UMICIS/242870, Marie Curie grants EUROTAST FP7-PEOPLE-2010 MC ITN, PALIMPSEST FP7-PEOPLE-2011-IEF 299101 and ORCA FP7-PEOPLE-2011-IOF 299075, a C2D2 Research Priming Fund grant partly funded by Wellcome Trust 097829, Swiss National Science Foundation grant 31003A-135688, the Novartis Foundation, the Novo Nordisk Foundation, the Max Planck Society and the University of York.

AUTHOR CONTRIBUTIONS

C.W. conceived the project, with input from M.J.C. R.S. and F.R. contributed samples. C.W., E.C., M.J.C., M.T.P.G., C.v.M., A.R. and Y.H. designed the experiments. C.W., E.C., N.S., C.T., A.R., Y.H., D.C.S.-G., S.C., S.F., H.U.L., P.N., C.D.K., J.V.O., K.Y.T. and E.E. performed the experiments. J.F.M.R., R.V., C.W., C.v.M., J.G., A.R., Y.H., R.Y.T., S.F., C.S., S.C., D.C.S.-G., J.H., J.A.S.C., L.H.H. and T.K. analyzed the data. S.B.-O., Y.H., E.W., C.M.L., M.T.P.G., M.J.C. and F.R. contributed material support to the project. Y.H. wrote the supplementary Raman section. C.W. wrote the manuscript, with critical input from C.M.L., M.T.P.G., M.J.C., C.v.M., E.W., E.C. and the remaining authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Marsh, P.D. Are dental diseases examples of ecological catastrophes? *Microbiology* **149**, 279–294 (2003).
2. Pihlstrom, B.L., Michalowicz, B.S. & Johnson, N.W. Periodontal diseases. *Lancet* **366**, 1809–1820 (2005).
3. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
4. Dewhirst, F.E. *et al.* The human oral microbiome. *J. Bacteriol.* **192**, 5002–5017 (2010).
5. Hujoel, P. Dietary carbohydrates and dental-systemic diseases. *J. Dent. Res.* **88**, 490–502 (2009).
6. Kuo, L.C., Polson, A.M. & Kang, T. Associations between periodontal diseases and systemic diseases: a review of the inter-relationships and interactions with diabetes, respiratory diseases, cardiovascular diseases and osteoporosis. *Public Health* **122**, 417–433 (2008).
7. Leishman, S.J., Do, H.L. & Ford, P.J. Cardiovascular disease and the role of oral bacteria. *J. Oral Microbiol.* **2**, 5781–5793 (2010).
8. Jin, Y. & Yip, H.K. Supragingival calculus: formation and control. *Crit. Rev. Oral Biol. Med.* **13**, 426–441 (2002).
9. Hardy, K. *et al.* Starch granules, dental calculus and new perspectives on ancient diet. *J. Archaeol. Sci.* **36**, 248–255 (2009).
10. Henry, A.G., Brooks, A.S. & Piperno, D.R. Microfossils in calculus demonstrate consumption of plants and cooked foods in Neanderthal diets (Shanidar III, Iraq; Spy I and II, Belgium). *Proc. Natl. Acad. Sci. USA* **108**, 486–491 (2011).
11. Adler, C.J. *et al.* Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat. Genet.* **45**, 450–455 (2013).
12. De La Fuente, C.P., Flores, S.V. & Moraga, M.L. DNA from human ancient bacteria: a novel source of genetic evidence from archaeological dental calculus. *Archaeometry* **55**, 767–778 (2013).
13. Linossier, A., Gajardo, M. & Olavarria, J. Paleomicrobiological study in dental calculus: *Streptococcus mutans*. *Scanning Microsc.* **10**, 1005–1013; discussion 1014 (1996).
14. Wang, J. *et al.* Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci. Rep.* **3**, 1843 (2013).
15. Fierer, N., Bradford, M.A. & Jackson, R.B. Toward an ecological classification of soil bacteria. *Ecology* **88**, 1354–1364 (2007).
16. Munro, C.L. & Grap, M.J. Oral health and care in the intensive care unit: state of the science. *Am. J. Crit. Care* **13**, 25–34 (2004).
17. Shay, K. Infectious complications of dental and periodontal diseases in the elderly population. *Clin. Infect. Dis.* **34**, 1215–1223 (2002).
18. Nakano, K. *et al.* Detection of oral bacteria in cardiovascular specimens. *Oral Microbiol. Immunol.* **24**, 64–68 (2009).
19. Gillespie, J.J. *et al.* PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect. Immun.* **79**, 4286–4298 (2011).
20. Willner, D. *et al.* Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc. Natl. Acad. Sci. USA* **108** (suppl. 1), 4547–4553 (2011).
21. Socransky, S.S. & Haffajee, A.D. Periodontal microbial ecology. *Periodontol. 2000* **38**, 135–187 (2005).
22. Marri, P.R. *et al.* Genome sequencing reveals widespread virulence gene exchange among human *Neisseria* species. *PLoS ONE* **5**, e11835 (2010).
23. Deguchi, T., Yasuda, M. & Ito, S. Management of pharyngeal gonorrhea is crucial to prevent the emergence and spread of antibiotic-resistant *Neisseria gonorrhoeae*. *Antimicrob. Agents Chemother.* **56**, 4039–4040 (2012).
24. Emonts, M., Hazelzet, J.A., de Groot, R. & Hermans, P.W. Host genetic determinants of *Neisseria meningitidis* infections. *Lancet Infect. Dis.* **3**, 565–577 (2003).
25. Goker, M. *et al.* Complete genome sequence of *Olsenella uli* type strain (VPI D76D-27C^{T}). *Stand. Genomic Sci.* **3**, 76–84 (2010).
26. Palmer, R.J. Composition and development of oral bacterial communities. *Periodontol. 2000* **64**, 20–39 (2014).
27. O'Brien-Simpson, N.M., Veith, P.D., Dashper, S.G. & Reynolds, E.C. Antigens of bacteria associated with periodontitis. *Periodontol. 2000* **35**, 101–134 (2004).
28. Amano, A., Nakagawa, I., Okahashi, N. & Hamada, N. Variations of *Porphyromonas gingivalis fimbriae* in relation to microbial pathogenesis. *J. Periodontal Res.* **39**, 136–142 (2004).
29. Sommer, M.O., Dantas, G. & Church, G.M. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* **325**, 1128–1131 (2009).
30. Xie, G. *et al.* Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Mol. Oral Microbiol.* **25**, 391–405 (2010).
31. D'Costa, V.M. *et al.* Antibiotic resistance is ancient. *Nature* **477**, 457–461 (2011).
32. Tanner, A.C. & Izard, J. *Tannerella forsythia*, a periodontal pathogen entering the genomic era. *Periodontol. 2000* **42**, 88–113 (2006).
33. Sharma, A. Virulence mechanisms of *Tannerella forsythia*. *Periodontol. 2000* **54**, 106–116 (2010).
34. Shimotahira, N. *et al.* The S-layer of *Tannerella forsythia* contributes to serum resistance and oral bacterial co-aggregation. *Infect. Immun.* **81**, 1198–1206 (2013).
35. Lee, S.W. *et al.* Identification and characterization of the genes encoding a unique surface (S-) layer of *Tannerella forsythia*. *Gene* **371**, 102–111 (2006).
36. Gorr, S.U. Antimicrobial peptides of the oral cavity. *Periodontol. 2000* **51**, 152–180 (2009).
37. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
38. Coxon, A., Tang, T. & Mayadas, T.N. Cytokine-activated endothelial cells delay neutrophil apoptosis *in vitro* and *in vivo*. A role for granulocyte/macrophage colony-stimulating factor. *J. Exp. Med.* **190**, 923–934 (1999).
39. Ryder, M.I. Comparison of neutrophil functions in aggressive and chronic periodontitis. *Periodontol. 2000* **53**, 124–137 (2010).
40. Brinkmann, V. & Zychlinsky, A. Neutrophil extracellular traps: is immunity the second function of chromatin? *J. Cell Biol.* **198**, 773–783 (2012).
41. Furugen, R., Hayashida, H. & Saito, T. *Porphyromonas gingivalis* and *Escherichia coli* lipopolysaccharide causes resistin release from neutrophils. *Oral Dis.* **19**, 479–483 (2013).
42. Kusminski, C.M., McTernan, P.G. & Kumar, S. Role of resistin in obesity, insulin resistance and Type II diabetes. *Clin. Sci.* (*Lond.*) **109**, 243–256 (2005).
43. Gracia-Arnaiz, M. Fat bodies and thin bodies. Cultural, biomedical and market discourses on obesity. *Appetite* **55**, 219–225 (2010).
44. Frei, R., Lauener, R.P., Crameri, R. & O'Mahony, L. Microbiota and dietary interactions: an update to the hygiene hypothesis? *Allergy* **67**, 451–461 (2012).
45. Palmer, S.A., Smith, O. & Allaby, R.G. The blossoming of plant archaeogenetics. *Ann. Anat.* **194**, 146–156 (2012).
46. Leonard, J.A. *et al.* Animal DNA in PCR reagents plagues ancient DNA research. *J. Archaeol. Sci.* **34**, 1361–1366 (2007).
47. Bocherens, H., Grupe, G., Mariotti, A. & Turban-Just, S. Molecular preservation and isotopy of Mesolithic human finds from the Ofnet cave (Bavaria, Germany). *Anthropol. Anz.* **55**, 121–129 (1997).
48. Oelze, V.M. *et al.* Multi-isotopic analysis reveals individual mobility and diet at the early iron age monumental tumulus of Magdalenenberg, Germany. *Am. J. Phys. Anthropol.* **148**, 406–421 (2012).
49. Oelze, V.M. *et al.* Early Neolithic diet and animal husbandry: stable isotope evidence from three Linearbandkeramik (LBK) sites in Central Germany. *J. Archaeol. Sci.* **38**, 270–279 (2011).
50. Schutkowski, H., Herrmann, B., Wiedemann, F., Bocherens, H. & Grupe, G. Diet, status and decomposition at Weingarten: trace element and isotope analyses on early mediaeval skeletal material. *J. Archaeol. Sci.* **26**, 675–685 (1999).
51. Turner-Walker, G. in *Advances in Human Paleopathology* (ed. Pinhasi, R. & Mays, S.) Ch. 1, 29 (John Wiley & Sons, New York, 2008).
52. Zijnge, V. *et al.* Oral biofilm architecture on natural teeth. *PloS One* **5**, e9321 (2010).
53. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–763 (2011).
54. Wood, J.W., Milner, G.R., Harpending, H.C. & Weiss, K.M. The osteological paradox—problems of inferring prehistoric health from skeletal samples. *Curr. Anthropol.* **33**, 343–370 (1992).
55. Meyer, F. *et al.* The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).
56. Vizcaíno, J.A. *et al.* The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **41**, D1063–D1069 (2013).

ONLINE METHODS

Study design and samples. Narrative and graphical overviews of the study design are provided in the **Supplementary Note** and **Supplementary Figure 17**. Archaeological material was obtained from the medieval St. Petri church and convent complex in Dalheim, Germany (**Supplementary Fig. 18**), and radiocarbon dated to c. 950–1200 CE (**Supplementary Table 5**). The assemblage was evaluated for pathologies (**Supplementary Table 6**), and dental tissues from four well-preserved adult skeletons (G12, B17, B61, B78) and two fauna (F1, F5) were selected for further analysis (**Supplementary Figs. 1**, **19** and **20**). Additionally, dental tissues from nine modern controls (P1–P5, P7, P8, P10, P13) with known dental health histories (**Supplementary Table 7**) were obtained under informed consent, and protocols were approved by the Zürich Ethics Commission (KEK ZH-Nr. 2012-0119).

Microscopy and spectroscopy. A mandibular incisor from B78 was sectioned longitudinally and examined according to standard protocols with a Tescan VEGA SEM using backscattered electron (BSE) imaging and EDS with a Si(Li) detector. Dental calculus deposits from B78 and P3 were fixed, decalcified and prepared into serial thin sections using modified standard protocols, followed by Gram and Hoechst staining and visualization using a Zeiss Axio Imager M2 and a Leica DMI6000 B microscope. Microfossils were obtained from dental calculus (G12, B17, B61, B78) and dental calculus/crown cementum (F5) deposits (**Supplementary Table 8**) using an incremental HCl decalcification protocol (**Supplementary Note**) and visualized using a Zeiss compound microscope under white and polarized light to identify pollen, phytoliths, starch granules and other debris (**Supplementary Table 9**) by comparison to reference collections. To evaluate mineralogical composition, Raman spectroscopy was applied to six calculus (G12, B17, B61, B78, P3, P13), nine dentine (G12, B17, B61, B78, P4, P5, P7, P8, P10) and five soil matrix (M1–M5) specimens using a HORIBA XploRA instrument (100× magnification and 532-nm laser wavelength) and analyzed for the main PO_4^{3-} peak position and peak area, as well as the peak intensity ratios of C-H (~2,940 cm^{-1}) *I*(CH) and main phosphate peak *I*(P) (**Supplementary Table 10**).

Isotope ratio mass spectrometry. Rib specimens from G12, B17, B61 and B78 were cleaned by abrasion, and collagen was extracted according to the method of Richards and Hedges[57] with an additional ultrafiltration step. Carbon and nitrogen isotopic values were measured in duplicate using a Thermo-Finnighan Delta XP continuous-flow isotope-ratio mass spectrometer following combustion in an elemental analyzer FLASH EA 2112 (**Supplementary Table 11**).

DNA extraction. Ancient samples were extracted in a dedicated ancient DNA laboratory at the University of Zürich Centre for Evolutionary Medicine in accordance with established contamination control precautions and workflows. DNA was extracted from dental calculus (G12, B17, B61, B78, P2), dentine (G12, B17, B61, B78), carious dentine (B17), abscessed alveolar bone (B78) and burial matrix (M1–M5) by phenol-chloroform extraction followed by Qiagen MinElute column purification (**Supplementary Tables 12–17**). Burial matrix and NaOCl-decontaminated dentine were tested for the presence of endogenous human DNA using targeted PCR and qPCR (**Supplementary Tables 18** and **19**). To optimize DNA extraction from dental calculus, five extraction buffers (A–E) and three decontamination methods were tested and compared. Two extraction buffers (A, 0.45 M EDTA, 10% proteinase K; B, 0.1 M EDTA, 10% proteinase K, 10 mM Tris-HCl, 10 mM NaCl, 2% SDS, 5 mM $CaCl_2$, 40 mM DTT) and three decontamination methods (2% NaOCl, 0.5 M EDTA wash, none) were selected for further analysis and were used in combination to produce nine DNA extracts from B61 and G12 dental calculus.

DNA library construction and sequencing. DNA extracts from the optimization experiment were built into nine shotgun libraries using a NEBNext Quick DNA Library Prep Master Mix Set (e6090) with DNA oligonucleotides containing a sample-specific multiplex index sequence (**Supplementary Fig. 21** and **Supplementary Table 20**). Libraries were amplified with Phusion HS II enzyme and sequenced on one lane of an Illumina HiSeq 2000 using single-end 1 × 100-bp chemistry, resulting in 93,677,545 reads after the removal of low-quality sequences (Illumina CASAVA 1.8.0, default settings, sequences <25 bp and/or with Phred scores <35 removed; **Supplementary Table 21**). Separately, 30 16S rRNA gene amplicon libraries were generated from dental calculus (G12, B17, B61, B78), dentine (G12, B17, B61, B78), carious dentine (B17) and alveolar bone abscess (B17) ancient DNA extracts generated using extraction buffer A and without previous decontamination. Universal primers targeting variable regions V3, V5 and V6 of the 16S rRNA gene were developed and tested *in silico* (**Supplementary Tables 22** and **23**) and *in vitro* (**Supplementary Fig. 22**). Each library was generated from a minimum of three amplifications (30–35 cycles) using Phusion HS II enzyme and 454 amplicon Fusion primers with multiplex identifiers (MIDs), and pooled 454 libraries were sequenced with a Roche GS Junior, resulting in 170,807 reads after the removal of low-quality sequences (Roche GS RunProcessor, default settings; **Supplementary Table 24**).

16S rRNA taxonomic classification. A reference data set containing full-length 16S ribosomal RNA sequences was constructed from the NCBI GenBank database, whereby all publicly available 16S ribosomal gene sequences found in the NCBI GenBank database were downloaded, screened for chimeras using uchime[58], aligned using the INFERNAL aligner[59] v1.0.2, trimmed and clustered at a sequence identity cutoff of 98% with a hierarchical clustering algorithm using sequence identity as the measure of distance and single linkage as the cluster metric. This data set has high overlap with both the Greengenes[60] (90%) and RDP[61] (92%) databases and was constructed to standardize filtering and alignment methods, as well as to streamline GenBank data retrieval for network analysis. Amplicon and shotgun sample reads were aligned to the reference OTU data set, and reads with a bit score of <40 or negative structure score were discarded. Sample reads were mapped to the reference OTUs by assigning the OTU ID of the most similar reference sequences. Conflicting OTU IDs were discarded. OTUs containing 16S rRNA gene sequences belonging to a reference genome or culture collection were assigned the consensus taxonomy of all such sequences in the OTU. In the case of OTUs that contained no reliable source of taxonomy, the taxonomy of the OTU was inferred by decreasing the clustering threshold until the point at which the OTU was merged with another in which sequences with reliable taxonomy existed.

Network analysis. Network analysis of community similarity was performed to compare the microbial communities of ancient dental samples to each other and to environmental samples deposited in GenBank and MG-RAST (project 128). Only environmental samples with at least 20 OTUs were considered (1,818 of 37,689), and only samples with at least 20% similarity to one of the ancient samples are shown in the network (315 of 1,818). The similarity between a pair of samples was calculated as the number of shared OTUs divided by the total number of different OTUs found in both samples. The network was rendered using the neato program from the Graphviz package (see URLs).

Phylogenetic tree. Ancient dental calculus, amplicon and shotgun OTU tables were merged, and a full-length 16S rRNA sequence representative for each OTU was chosen. Phylogenetic relationships were inferred with FastTree[62] v2.1.3 (generalized time-reversible model).

Validation of results using the RDP and QIIME pipelines. To confirm that the taxonomic characterization of ancient dental samples was robust to database choice and clustering parameters, the 16S rRNA amplicon data were reanalyzed using the Greengenes database (v.4Feb2011) and the RDP Pyrosequencing[61] and QIIME[63] pipelines. Only reads of ≥70 bp with 100% identity to both forward and reverse primers were analyzed. OTUs were clustered at 97% identity, and singleton OTUs were discarded (**Supplementary Table 25**). The OTU table was rarefied to 1,265 sequences/sample and analyzed at the L2, L5 and L6 levels. Alpha and beta diversity were calculated using QIIME default parameters. The BIOM file for these data is available as **Supplementary Data Set 1**. This OTU table was merged with an OTU table generated from the HMP data set using the same parameters, and the two data sets were compared using Principal Coordinates Analysis; the BIOM file for these data is available as **Supplementary Data Set 2**.

Source tracking. To test for contamination in the ancient dental samples, Bayesian microbial source tracking[53] was performed (1,000 'burn-in' iterations

using Gibbs sampling with 25 random restarts) on the merged OTU file using HMP plaque, HMP skin, HMP gut and ancient tooth root (environmental proxy) as sources.

Dietary DNA analysis. Shotgun reads ≥75 bp in length were searched against a complete collection of full mitochondrial and chloroplast genome sequences published as of July 2012 (>6,000 organelle genomes) using BLASTN. Results were accepted only if they exhibited 100% query coverage and 100% sequence identity, were not hits to 16S or 23S rRNA genes and did not match more than one genus perfectly, and any secondary hits outside the genus of the first hit had to show at least two diagnostic point mutations relative to the perfect hit.

Total taxonomic characterization of dental calculus. Library reads were pooled by individual (B61, S1–S4; G12, S5–S8) and *de novo* assembled into 2,005,273 contigs using Velvet[64] v.1.02.3 (*k*-mer length of 29 bp, minimum of 100-bp contig length) (**Supplementary Table 26**). Contigs were searched against the NCBI nr and gss databases available as of July 2012 using Megablast, filtered for highly unique, high-scoring top hits (>95-bp alignment, >97% identity, *E* value of $<1 \times 10^{-14}$). A total of 61,584 contigs passing these filters were assigned taxonomy.

Pathogen analysis. Contigs were further filtered to remove contigs with second hits of comparable quality and >90% identity to other taxa, resulting in 53,924 highly unique contigs that can be reasonably assigned to a single species. Species-level assignments were then cross-referenced against the PATRIC database[19], resulting in 40 putative pathogen identifications. To determine whether these species assignments were reasonable for the oral cavity, we applied the same BLAST and conditional filter approach to shotgun metagenomic contigs reported for 109 HMP supragingival dental plaque samples and compared the results. Feature information for each ancient contig was retrieved from the top-hit BLAST results and manually screened for putative genes associated with virulence, drug resistance, plasmids, transposons and phages with annotations in PubMed records.

Antibiotic resistance analysis. Sequences for all identified taxa were screened for putative antibiotic resistance elements using three methods: (i) BLASTX search against the Antibiotic Resistance Database (ARDB)[65], (ii) BLASTX search against the NCBI nr database followed by keyword search of translated gene function and (iii) manual search of gene annotations assigned to pathogens.

Genome reconstruction. All G12 contigs of ≥100 bp were searched against the NCBI nt and gss databases using Megablast and filtered for contigs aligning to *T. forsythia* strain ATCC 43037 with an *E* value of $\leq 1 \times 10^{-6}$ within the top 100 hits. Filtered contigs were pooled and submitted to the BLAST Ring Image Generator (BRIG)[66] tool for mapping. Using BRIG, contigs were aligned to *T. forsythia* strain ATCC 43037 using the Megablast search option and a sequence identity cutoff of ≥95%. In cases where a contig aligned to the *T. forsythia* genome more than once, the alignment with the highest bit score was mapped. In cases where multiple alignments with identical top bit scores were observed, the contig was mapped to all top bit score loci, but the depth of coverage for each locus was divided by the number of loci. Genes not mapped in the assembly and large gaps (**Supplementary Fig. 23**) were analyzed for function.

Protein analysis. Total proteins were extracted from dental calculus (G12, B17, B61, B78, P1, P2), dentine (G12, B17, B61, B78), carious dentine (B17), abscessed alveolar bone (B78) and dental calculus/crown cementum (F1, F5) and from four negative extraction controls using a modified filter-aided sample preparation (FASP)[67] protocol. A total of 290,466 MS/MS spectra were generated using 3 instruments (LTQ-Orbitrap Velos, Q-Exactive Hybrid Quadrupole Oritrap and MaXis UHR-Qq-TOF) (**Supplementary Table 27**). Tandem mass spectra were converted to Mascot generic format using ProteoWizard v.2.2.3101 with vendor peak picking option for MS level 2 and deisotoped and deconvoluted using the H-Score script[68]. ProteinPilot v.4 was used to analyze protein modification and damage patterns (**Supplementary Table 28**). MS/MS peak lists were searched using Mascot v.2.3.02 against all proteins in UniProtKB/SwissProt as of 31 October 2012 and two custom protein databases built from the Human Oral Microbiome Database (HOMD)[69] as of 11 October 2012 and all complete soil bacterial genomes in GenBank as of 22 February 2012. The results were further validated using Scaffold v.4.0.5, resulting in 12,609 unique peptide identifications resolving to 589 proteins identified with >99% confidence and ≥2 unique peptides. Contaminants were identified and removed (**Supplementary Table 29**). Metadata for human proteins was retrieved using the GeneCards v.3 GeneALaCart tool[70] and used to manually classify each protein into six categories: innate immune system, adaptive immune system, blood coagulation, digestion, structure and support, and other. Protein interaction and gene ontology (GO) information was obtained using STRING 9.0 (ref. 37) in protein mode. Bacterial proteins were binned by length (group 1, <15 residues; group 2, >15 residues) and searched against the NCBI database using BLASTP (group 1, expect value 20,000, PAM30 Score Matrix; group 2, expect value 1,000, BLOSUM62 Score Matrix). Resulting BLASTP files were then parsed using MEGAN[71] and analyzed for protein function using SEED hierarchy[72].

57. Richards, M.P. & Hedges, R.E.M. Stable isotope evidence for similarities in the types of marine foods used by late mesolithic humans at sites along the Atlantic coast of Europe. *J. Archaeol. Sci.* **26**, 717–722 (1999).
58. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).
59. Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
60. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).
61. Cole, J.R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141–D145 (2009).
62. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
63. Kuczynski, J. *et al.* Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr. Protoc. Bioinformatics* Chapter 10, Unit 10.17 (2011).
64. Zerbino, D.R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
65. Liu, B. & Pop, M. ARDB antibiotic resistance genes database. *Nucleic Acids Res.* **37**, D443–D447 (2009).
66. Alikhan, N.F., Petty, N.K., Ben Zakour, N.L. & Beatson, S.A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).
67. Cappellini, E. *et al.* Resolution of the type material of the Asian elephant, *Elephas maximus* Linnaeus, 1758 (Proboscidea, Elephantidae). *Zool. J. Linn. Soc.* **170**, 222–232 (2014).
68. Savitski, M.M., Mathieson, T., Becher, I. & Bantscheff, M. H-score, a mass accuracy driven rescoring approach for improved peptide identification in modification rich samples. *J. Proteome Res.* **9**, 5511–5516 (2010).
69. Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database* (*Oxford*) **2010**, baq013 (2010).
70. Stelzer, G. *et al. In-silico* human genomics with GeneCards. *Hum. Genomics* **5**, 709–717 (2011).
71. Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
72. Mitra, S. *et al.* Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics* **12** (suppl. 1), S21 (2011).

Chapter 6

Other contributions

Microbiota-Derived Hydrogen Fuels Salmonella Typhimurium Invasion of the Gut Ecosystem
This study was led by Dr. Maier to investigate the process of gut ecosystem colonization by S. Typhimurium during an infection episode. Knockout studies in the S. Tm SL1344 showed that a non-functional copy of the Fe-Ni uptake-type hydrogenase operon (which facilitates the consumption of hydrogen) conferred fitness disadvantage to the mutant mice as compared to wild type S. Tm in the gut with an unperturbed microbiota. However, the fitness disadvantage disappeared in a low complexity microbiota mice (or mice that did not harbor a healthy microbiota, germ free mice, streptomycin-pre treated mice).Furthermore, S. Tm growth also suffered when other hydrogen consuming bacteria were introduced in the gut. Since hydrogen is not produced or consumed by the host (mice, human), hydrogen derived from the gut microbiota seems to boost the growth of S. Tm. As a part of this project, I was responsible for identifying genes responsible for hydrogen producing and consuming enzymes present in the gut metagenome of six different species. This was with the aim to assess the availability of hydrogen and its metabolism in different guts.

IFN-γ hinders recovery from mucosal inflammation during antibiotic therapy of Salmonella gut infection
This study was led by Dr. Dolowschiak to characterize an episode of S .Tm infection and recovery in mice. The mice were initially orally infected and then subsequently treated with a standard antibiotic to promote recovery. Thereafter several inflammatory markers were measured at routine intervals using RT qPCR. The measurements were made on 52 genes including a variety of cytokines, chemo-attractants, granulocyte infiltration, antimicrobials, interleukins and housekeeping genes. For each day, measurements were taken from three mice. I was responsible for the data analysis focusing on the shift of the expression patterns over the course of infection to confirm whether these inflammatory markers returned to their baseline values after the infection.

Chapter 7

Concluding Remarks

A major contribution of this work has been high resolution characterization of microbial communities associated to two different anatomical sites. By using the methods presented in this work, it is possible to supplement or partially replace traditional culture methods in favor of whole genome shotgun sequencing approaches which have the power to identify nearly all species present in a clinical sample with high taxonomic resolution. These methods rely solely on sequencing with minimal experimental manipulation, streamlining the detection process. We determine that strain typing followed by near complete whole genome reconstruction is feasible using whole metagenome shotgun sequencing data for dominant, over-represented microbes using total DNA extracted from CF sputum samples.

This technique essentially builds upon the knowledge of existing comprehensive multilocus sequence typing (MLST) databases of known CF pathogens to speed up strain identification. The main advantage of this method is that presence of multiple strains can be detected for a number of different pathogens with relative ease from routinely collected CF sputum samples without adding elaborate experimental work. Thus temporal changes can be tracked in strain diversity to prepare comprehensive histories of infection episodes. This knowledge can provide a detailed view of the changes that take place a microbial community during the course of infection, promoting deeper understanding of the infection dynamics. However, the reliance of this approach upon well populated MLST databases places constraints on the species which can be identified. Specifically, it is only possible to carry out strain identification for species that have been previously subjected to MLST. Thus strain typing using this genomic technique is contingent upon availability of well populated MLST databases. Nonetheless, as more and more species undergo experimental strain typing we anticipate an increasing relevance of strictly genomic analysis-based techniques in a clinical setting in the near future.

The MLST data can be further complemented by assembling whole genomes to learn about possible genome rearrangements and gain of any ge-

netic elements in the newly assembled strain. This information can be further correlated with the strain identity to functionally characterize the strains. Using this approach, a previously documented MLST strain of a known opportunistic pathogen was detected in the sputum sample of one CF patient and it was further characterized by assembling its near complete genome. This led to the detection of a virulence associated secretion system (T6SS) which had previously not been widely observed in this bacterial species. Knowledge of such virulence linked adaptations have the potential to impact the course of treatment and is thus clinically relevant information. Gathering data about the presence of different strain types and their virulence potential over time can help in design of more informed treatments. Additionally, one can distinguish between infections caused by a singular strain from those caused by a cloud of strains and also monitor whether any of the strains can outcompete others to establish clonal infections. This information is particularly relevant in case of CF patients since CF pathogens such as *p. aeruginosa* are known to undergo mutations that enables their environmental, virulent strains (responsible for causing the infection) to adapt and thrive in the lung environment. Thus whole or partial genome reconstruction provides a unique opportunity to build catalogues of such mutations observed across gene sets of known function and identify evolutionary hotspots in the pathogen genome without routinely engaging in experimental techniques. This information is crucial to developing a thorough understanding of pathogen evolution over the course of infection.

The sequence based analysis techniques discussed in this work represents a powerful addition to a clinicians toolbox particularly for characterizing microbial infection. Emerging adaptations such as antibiotic resistance can be detected by scanning the entire community's metagenome against the existing antibiotic resistance databases. Since clinical testing for antibiotic resistance is limited to bacteria that can be cultivated in laboratories, it does not account for resistances harbored by the uncultivable fraction of the microbial community. Thus sequence based prediction of resistances is likely to present a more comprehensive account as compared to current culture based methods. Moreover, since sequence-based prediction of antibiotic resistance relies on direct observation of the resistance conferring gene sets or relevant mutation, it may be possible to detect such adaptations before they translate into an observable phenotype that can be detected in clinical resistance testing (i.e. before the resistance gene or mutation fixates in a microbial population). Such rapidity may prove useful for aggressive infections for which time is of the essence. However, sequence-based prediction of antibiotic resistance requires high quality, publically available datasets that have been experimentally confirmed. The development of such databases is currently restricted by limited understanding of the resistance conferring mechanisms. As these databases become more comprehensive, genomic prediction of antibiotic resistance will find greater application

in assisting treatment design.

Next, to investigate the evolution of human associated microbial communities, we examined a calcified form of an oral biofilm (dental plaque) known as dental calculus. We used a combination of non-targeted and 16S WGS sequencing to assess whether dental calculus contains meaningful information about the ancient oral microbiome. One of the most important findings of this study was that ancient dental calculus is a rich source of information about the ancient oral microbial community and contians little contamination from the environmental microbes. This finding is likely going to impact the way archaeologists currently investigate ancient human lifestyle. It is one of the first studies to establish that dental calculus contains diagnostically useful DNA to identify members of the ancient oral microbiota, opportunistic upper respiratory pathogens and dietary information. This study raises important questions about the evolution of microbes and the nature of relationship with their human host. Using calcified plaque samples from different time periods, it is potentially possible to investigate whether oral microbial community has undergone major rearrangements in terms of its composition. Another interesting question to investigate would be the evolution of oral pathogens. In this study, we reconstruct a near complete genome of a periodontal pathogen and compare it to its contemporary counterpart and identify lack of pathogenicity associated genomic islands. Since dental calculus preserves well over extended periods of time, it can serve as an excellent source to investigate ancient human health and its associated microbiota from different eras. Cross comparison with contemporary knowledge can also bear important implications on improving management of human health.

So far we have discussed the advantages of WGS sequencing for characterizing human associated microbial populations. Indeed, WGS provides a promising alternative to the current culture based techniques, however it also presents a number of shortcomings. Firstly, the untargeted nature of WGS can lead to a potential bias for microbes with differing genome sizes such as bacteria and fungi. Furthermore, due to their extremely large size, eukaryotic genomes require greater sequencing depth before they can be assembled compared to bacterial genomes. Thus eukaryotic microbes are more likely to be underrepresented in assembled data. This shortcoming can be addressed by either increasing the sequencing depth or by using enrichment protocols to capture the eukaryotic population in combination with WGS sequencing. Another shortcoming is that current WGS data analysis tools strongly depend on availability of well curated genomic databases. This leads to annotation of only a fraction of the sequenced data at varying levels of resolution. To improve the rate of annotation, further work is needed to build comprehensive, high-resolution databases or development of more sophisticated binning techniques that are able to differentiate between genomes based on their intrinsic properties. In chapter 4, we use a combination of GC content, contig entropy and length to

facilitate the detection of highly abundant clonal species without using homology based searches. This was carried out with the aim of detecting previously unknown, abundant microbes. While there were no new species detected in the CF patients, this approach has the potential to discriminate dominant, clonal colonizers, even previously unobserved ones and is widely applicable to samples obtained not only from clinics, but also for other microbe-rich environments as well.

The work discussed in this dissertation has been centered around snapshot data that provides clinically relevant information and raises interesting questions about the changes that follow in human associated microbial communities during the course of an infection. Future work would be to apply the genomic methods described here to build temporal histories and characterize the community dynamics in patient and healthy lungs. By determining that total DNA from non-invasive samples such as sputum is a rich source of information about the endogenous lung microbial community, we hope to facilitate high resolution routine monitoring of the lung microbial community.

Detection of a possibly new, yet unnamed *Achromobacter* species in one CF patient highlights the need to combine experimental and genomic techniques to identify diagnostically useful species markers to allow accurate identification of the infecting pathogen. To date, there has been no clear consensus on the effect of *Achromobacter* infection on prognosis of CF. Accurate identification of the infecting *Achromobacter* species is likely going to help clear the ambiguity on the impact of infecting species on lung function and disease prognosis.

In this study, while most CF patients showed visibly reduced microbial diversity (as expected in a diseased lung), one advanced CF patient harbored a rich and diverse lung microbial community along with a known CF pathogen. Temporal studies are required to understand the effect of a dominant colonizer on the community structure and composition. It would be especially interesting to investigate nature of interaction between various pathogens and other members of lung microbial community. This can help understand the distinct changes that occur in the lung microbiota in the presence of different pathogens.

We hope that one of the long term desirable outcomes of such studies would be the development of microbiome based therapies to restore the healthy commensal population and help in management (if not treatment) of complex, polymicrobial infections.

Part III

Back Matter

Bibliography

[1] Andrew Adey, Hilary G Morrison, Xu Xun, Jacob O Kitzman, Emily H Turner, Bethany Stackhouse, Alexandra P MacKenzie, Nicholas C Caruccio, Xiuqing Zhang, Jay Shendure, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology*, 11(12):1, 2010.

[2] Bernhard Hube Amy Whittington, Neil A.R. Gow. From commensal to pathogen: Candida albicans. *Human Fungal Pathogens, 2nd Edition*, 1(1):3–7, 2014.

[3] M Avila, D M Ojcius, and Ö Yilmaz. The oral microbiota: living with a permanent guest. *DNA and cell biology*, 2009.

[4] Amir Azarpazhooh and James L Leake. Systematic review of the association between respiratory diseases and oral health. *Journal of periodontology*, 77(9):1465–1482, 2006.

[5] Fredrik Bäckhed, Claire M Fraser, Yehuda Ringel, Mary Ellen Sanders, R Balfour Sartor, Philip M Sherman, James Versalovic, Vincent Young, and B Brett Finlay. Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell Host and Microbe*, 12(5):611–622, November 2012.

[6] Fredrik Bäckhed, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon. Host-bacterial mutualism in the human intestine. *science*, 307(5717):1915–1920, 2005.

[7] A Bankevich, S Nurk, and D Antipov. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of . . .*, 2012.

[8] Teresa M Barbosa and Stuart B Levy. The impact of antibiotic use on resistance development and persistence. *Drug resistance updates*, 3(5):303–311, 2000.

[9] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon,

Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004.

[10] Robert P Baughman, Joseph E Thorpe, Joseph Staneck, Mitchell Rashkin, and Peter T Frame. Use of the protected specimen brush in patients with endotracheal or tracheostomy tubes. *Chest*, 91(2):233–236, 1987.

[11] Clifford J Beall, Alisha G Campbell, Daniel M Dayeh, Ann L Griffen, Mircea Podar, and Eugene J Leys. Single Cell Genomics of Uncultured, Health-Associated Tannerella BU063 (Oral Taxon 286) and Comparison to the Closely Related Pathogen Tannerella forsythia . *PLoS ONE*, 9(2):e89398, February 2014.

[12] J M Beck, V B Young, and G B Huffnagle. The microbiome of the lung. *Translational Research*, 2012.

[13] S J Bent, J D Pierson, L J Forney, R Danovaro, G M Luna, A Dell'Anno, and B Pietrangeli. Measuring Species Richness Based on Microbial Community Fingerprints: the Emperor Has No Clothes. *Applied and Environmental Microbiology*, 73(7):2399–2401, April 2007.

[14] Elena Biagi, Lotta Nylund, Marco Candela, Rita Ostan, Laura Bucci, Elisa Pini, Janne Nikkïla, Daniela Monti, Reetta Satokari, Claudio Franceschi, et al. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PloS one*, 5(5):e10667, 2010.

[15] Surinder S Birring, Sergio Matos, Ronnak B Patel, Benjamin Prudon, David H Evans, and Ian D Pavord. Cough frequency, cough sensitivity and health status in patients with chronic cough. *Respiratory medicine*, 100(6):1105–1109, 2006.

[16] John M Boyce and Nancy L Havill. Nosocomial antibiotic-associated diarrhea associated with enterotoxin-producing strains of methicillin-resistant staphylococcus aureus. *American Journal of Gastroenterology*, 100(8):1828–1834, 2005.

[17] Michael Brudno, Alexander Poliakov, Simon Minovitsky, Igor Ratnere, and Inna Dubchak. Multiple whole genome alignments and novel biomedical applications at the vista portal. *Nucleic acids research*, 35(suppl 2):W669–W674, 2007.

[18] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC bioinformatics*, 10(1):1, 2009.

[19] Emily S Charlson, Kyle Bittinger, Andrew R Haas, Ayannah S Fitzgerald, Ian Frank, Anjana Yadav, Frederic D Bushman, and Ronald G Collman. Topographical continuity of bacterial populations in the healthy human respiratory tract. *American journal of respiratory and critical care medicine*, 184(8):957–963, 2011.

[20] J C Clemente, E C Pehrsson, and M J Blaser. The microbiome of uncontacted Amerindians. *Science*, 2015.

[21] Human Microbiome Project Consortium et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[22] Elizabeth K Costello, Christian L Lauber, Micah Hamady, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009.

[23] Elizabeth K Costello, Christian L Lauber, Micah Hamady, Noah Fierer, Jeffrey I Gordon, and Rob Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, December 2009.

[24] Louise Cullen and Siobhán McClean. Bacterial adaptation during chronic respiratory infections. *Pathogens*, 4(1):66–89, 2015.

[25] R M da Silva, D A Caugant, and R Josefsen. Characterization of Streptococcus constellatus strains recovered from a brain abscess and periodontal pockets in an immunocompromised patient. *Journal of . . .*, 2004.

[26] Gautam Dantas, Morten OA Sommer, Rantimi D Oluwasegun, and George M Church. Bacteria subsisting on antibiotics. *Science*, 320(5872):100–103, 2008.

[27] Aaron E Darling, Bob Mau, and Nicole T Perna. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE*, 5(6):e11147, 2010.

[28] Richard de Boer, Remco Peters, Sonja Gierveld, Tim Schuurman, Mirjam Kooistra-Smid, and Paul Savelkoul. Improved detection of microbial DNA after bead-beating before DNA isolation. *Journal of microbiological methods*, 80(2):209–211, February 2010.

[29] T de Sablet and C Chassard. Human microbiota-secreted factors inhibit shiga toxin synthesis by enterohemorrhagic Escherichia coli O157: H7. *Infection and . . .*, 2009.

[30] Tom O Delmont, Patrick Robe, Ian Clark, Pascal Simonet, and Timothy M Vogel. Metagenomic comparison of direct and indirect soil dna extraction approaches. *Journal of microbiological methods*, 86(3):397–400, 2011.

[31] Les Dethlefsen and David A Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4554–4561, 2011.

[32] Floyd E Dewhirst, Tuste Chen, Jacques Izard, Bruce J Paster, Anne C R Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G Wade. The Human Oral Microbiome. *Journal of Bacteriology*, 192(19):5002–5017, September 2010.

[33] Floyd E Dewhirst, Tuste Chen, Jacques Izard, Bruce J Paster, Anne CR Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G Wade. The human oral microbiome. *Journal of bacteriology*, 192(19):5002–5017, 2010.

[34] Martha L Diaz-Torres, Aurelie Villedieu, Nigel Hunt, Rod McNab, David A Spratt, Elaine Allan, Peter Mullany, and Michael Wilson. Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. *FEMS Microbiology Letters*, 258(2):257–62, May 2006.

[35] Robert P Dickson, John R Erb-Downward, Christine M Freeman, Lisa McCloskey, James M Beck, Gary B Huffnagle, and Jeffrey L Curtis. Spatial variation in the healthy human lung microbiome and the adapted island model of lung biogeography. *Annals of the American Thoracic Society*, 12(6):821–830, 2015.

[36] Robert P Dickson, John R Erb-Downward, and Gary B Huffnagle. Towards an ecology of the lung: new conceptual models of pulmonary microbiology and pneumonia pathogenesis. *The lancet Respiratory medicine*, 2(3):238–246, 2014.

[37] Robert P Dickson, John R Erb-Downward, Fernando J Martinez, and Gary B Huffnagle. The microbiome and the respiratory tract. *Annual review of physiology*, 78:481–504, 2016.

[38] Maria G Dominguez-Bello, Elizabeth K Costello, Monica Contreras, Magda Magris, Glida Hidalgo, Noah Fierer, and Rob Knight. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, 107(26):11971–11975, 2010.

[39] Christine Dominianni, Rashmi Sinha, James J Goedert, Zhiheng Pei, Liying Yang, Richard B Hayes, and Jiyoung Ahn. Sex, body mass index, and dietary fiber intake influence the human gut microbiome. *PLoS ONE*, 10(4):e0124599, 2015.

[40] Markus J Ege, Melanie Mayer, Anne-Cécile Normand, Jon Genuneit, William OCM Cookson, Charlotte Braun-Fahrländer, Dick Heederik, Renaud Piarroux, and Erika von Mutius. Exposure to environmental microorganisms and childhood asthma. *New England Journal of Medicine*, 364(8):701–709, 2011.

[41] Joan G Ehrenfeld, Beth Ravit, and Kenneth Elgersma. Feedback in the plant-soil system. *Annu. Rev. Environ. Resour.*, 30:75–115, 2005.

[42] S Dusko Ehrlich, MetaHIT Consortium, et al. Metahit: The european union project on metagenomics of the human intestinal tract. In *Metagenomics of the human body*, pages 307–316. Springer, 2011.

[43] Jonathan A Eisen. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol*, 5(3):e82, 2007.

[44] John R Erb-Downward, Deborah L Thompson, Meilan K Han, Christine M Freeman, Lisa McCloskey, Lindsay A Schmidt, Vincent B Young, Galen B Toews, Jeffrey L Curtis, Baskaran Sundaram, et al. Analysis of the lung microbiome in the healthy smoker and in copd. *PloS one*, 6(2):e16384, 2011.

[45] Wenguang Fan, Guicheng Huo, Xiaomin Li, Lijie Yang, and Cuicui Duan. Impact of diet in shaping gut microbiota revealed by a comparative study in infants during the first six months of life. *J. Microbiol. Biotechnol*, 24(2):133–143, 2014.

[46] Milan Fedurco, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research*, 34(3):e22, 2006.

[47] G R Feehery, E Yigit, S O Oyola, and B W Langhorst. A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS ONE*, 2013.

[48] Laura M Filkins and George A OToole. Cystic fibrosis lung infections: polymicrobial, complex, and hard to treat. *PLoS Pathog*, 11(12):e1005258, 2015.

[49] S Filoche, L Wong, and C H Sissons. Oral biofilms: emerging concepts in microbial ecology. *Journal of dental research*, 2010.

[50] GEORGE E FOX, Kenneth R Pechman, and Carl R Woese. Comparative cataloging of 16s ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *International Journal of Systematic and Evolutionary Microbiology*, 27(1):44–57, 1977.

[51] Rima B Franklin, Douglas R Taylor, and Aaron L Mills. Characterization of microbial communities using randomly amplified polymorphic dna (rapd). *Journal of Microbiological Methods*, 35(3):225–235, 1999.

[52] AG Fredrickson and Gregory Stephanopoulos. Microbial competition. *Science*, 213(4511):972–979, 1981.

[53] Steven R Gill, Mihai Pop, Robert T DeBoy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *science*, 312(5778):1355–1359, 2006.

[54] E M Glass, J Wilkening, and A Wilke. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harbor . . .*, 2010.

[55] R N Glud, F Wenzhöfer, M Middelboe, and K Oguri. High rates of microbial carbon turnover in sediments in the deepest oceanic trench on Earth. *Nature*, 2013.

[56] Armin J Grau, Heiko Becher, Christoph M Ziegler, Christoph Lichy, Florian Buggle, Claudia Kaiser, Rainer Lutz, Stefan Bültmann, Michael Preusch, and Christof E Dörfer. Periodontal disease as a risk factor for ischemic stroke. *Stroke; a journal of cerebral circulation*, 35(2):496–501, February 2004.

[57] George Hajishengallis and Richard J Lamont. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (psd) model of periodontal disease etiology. *Molecular oral microbiology*, 27(6):409–419, 2012.

[58] Marie A Hildebrandt, Christian Hoffmann, Scott A Sherrill-Mix, Sue A Keilbaugh, Micah Hamady, Ying-Yu Chen, Rob Knight, Rexford S Ahima, Frederic Bushman, and Gary D Wu. High-fat diet determines the composition of the murine gut microbiome independently of obesity. *Gastroenterology*, 137(5):1716–24.e1–2, November 2009.

[59] JD Hillman, AL Dzuback, and SW Andrews. Colonization of the human oral cavity by a streptococcus mutans mutant producing increased bacteriocin. *Journal of Dental Research*, 66(6):1092–1094, 1987.

[60] Kenya Honda and Dan R Littman. The microbiome in infectious disease and inflammation. *Annual review of immunology*, 30:759, 2012.

[61] Edmond Y Huang, Takuya Inoue, Vanessa A Leone, Sushila Dalal, Ketrija Touw, Yunwei Wang, Mark W Musch, Betty Theriault, Kazuhide Higuchi, Sharon Donovan, Jack Gilbert, and Eugene B Chang. Using corticosteroids to reshape the gut microbiome: implications for inflammatory bowel diseases. *Inflammatory bowel diseases*, 21(5):963–972, May 2015.

[62] Philip Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome biology*, 3(2):1, 2002.

[63] Howard F Jenkinson and Richard J Lamont. Oral microbial communities in sickness and in health. *Trends in Microbiology*, 13(12):589–595, December 2005.

[64] Lars Juhl Jensen, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. eggnog: automated construction and annotation of orthologous groups of genes. *Nucleic acids research*, 36(suppl 1):D250–D254, 2008.

[65] Shari A Jones, Mathias Jorgensen, Fatema Z Chowdhury, Rosalie Rodgers, James Hartline, Mary P Leatham, Carsten Struve, Karen A Krogfelt, Paul S Cohen, and Tyrrell Conway. Glycogen and maltose utilization by escherichia coli o157: H7 in the mouse intestine. *Infection and immunity*, 76(6):2531–2540, 2008.

[66] Jesse Stombaugh William Anton Walters Antonio González J Gregory Caporaso Rob Knight Justin Kuczynski. Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, CHAPTER:Unit10.7, December 2011.

[67] Nobuhiko Kamada, Grace Y Chen, Naohiro Inohara, and Gabriel Núñez. Control of pathogens and pathobionts by the gut microbiota. *Nature immunology*, 14(7):685–690, 2013.

[68] M Kanehisa. The KEGG database. *silico simulation of biological processes*, 2002.

[69] D Y Kil and K S Swanson. Companion animals symposium: role of microbes in canine and feline health. *Journal of animal science*, 89(5):1498–1505, May 2011.

[70] Jennifer L Kirk, Lee A Beaudette, Miranda Hart, Peter Moutoglis, John N Klironomos, Hung Lee, and Jack T Trevors. Methods of studying soil microbial diversity. *Journal of microbiological methods*, 58(2):169–188, 2004.

[71] Jeremy E Koenig, Aymé Spor, Nicholas Scalfone, Ashwana D Fricker, Jesse Stombaugh, Rob Knight, Largus T Angenent, and Ruth E Ley. Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4578–4585, 2011.

[72] David J Lane, Bernadette Pace, Gary J Olsen, David A Stahl, Mitchell L Sogin, and Norman R Pace. Rapid determination of 16s ribosomal rna sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, 82(20):6955–6959, 1985.

[73] George A O'Toole Laura M Filkins. Cystic Fibrosis Lung Infections: Polymicrobial, Complex, and Hard to Treat. pages 1–8, December 2015.

[74] Jean Guy LeBlanc, Christian Milani, Graciela Savoy de Giori, Fernando Sesma, Douwe Van Sinderen, and Marco Ventura. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Current opinion in biotechnology*, 24(2):160–168, 2013.

[75] Martina I Lefterova, Carlos J Suarez, Niaz Banaei, and Benjamin A Pinsky. Next-generation sequencing for infectious disease diagnosis and management: a report of the association for molecular pathology. *The Journal of Molecular Diagnostics*, 17(6):623–634, 2015.

[76] M J Levene. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science*, 299(5607):682–686, January 2003.

[77] Ruth E Ley, Micah Hamady, Catherine Lozupone, Peter J Turnbaugh, Rob Roy Ramey, J Stephen Bircher, Michael L Schlegel, Tammy A Tucker, Mark D Schrenzel, Rob Knight, et al. Evolution of mammals and their gut microbes. *Science*, 320(5883):1647–1651, 2008.

[78] D Li, R Luo, C M Liu, C M Leung, H F Ting, and K Sadakane. MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 2016.

[79] Yan Wei Lim, Robert Schmieder, Matthew Haynes, Dana Willner, Mike Furlan, Merry Youle, Katelynn Abbott, Robert Edwards, Jose Evangelista, Douglas Conrad, and Forest Rohwer. Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *Journal of Cystic Fibrosis*, 12(2):154–164, March 2013.

[80] Li Liu, Xiaowei Chen, Geir Skogerbø, Peng Zhang, Runsheng Chen, Shunmin He, and Da-Wei Huang. The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics*, 100(5):265–270, 2012.

[81] C Y Low and C Rotstein. Emerging fungal infections in immunocompromised patients. *F1000 Med Rep*, 2011.

[82] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, et al. Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):1, 2012.

[83] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[84] D Mariat, O Firmesse, F Levenez, VD Guimarães, H Sokol, J Dore, G Corthier, and JP Furet. The firmicutes/bacteroidetes ratio of the human microbiota changes with age. *BMC microbiology*, 9(1):1, 2009.

[85] Victor M Markowitz, Natalia N Ivanova, Ernest Szeto, Krishna Palaniappan, Ken Chu, Daniel Dalevi, I-Min A Chen, Yuri Grechkin, Inna Dubchak, Iain Anderson, et al. Img/m: a data management and analysis system for metagenomes. *Nucleic acids research*, 36(suppl 1):D534–D538, 2008.

[86] P D Marsh. Are dental diseases examples of ecological catastrophes? *Microbiology*, 2003.

[87] K J Mattila, V V Valtonen, M S Nieminen, and S Asikainen. Role of infection as a risk factor for atherosclerosis, myocardial infarction, and stroke. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 26(3):719–734, March 1998.

[88] M B Miller and B L Bassler. Quorum sensing in bacteria. *Annual Reviews in Microbiology*, 2001.

[89] Shirajum Monira, Shota Nakamura, Kazuyoshi Gotoh, Kaori Izutsu, Haruo Watanabe, Nur Haque Alam, Hubert Ph Endtz, Alejandro Cravioto, Sk Ali, Takaaki Nakaya, et al. Gut microbiota of healthy and malnourished children in bangladesh. *Frontiers in microbiology*, 2:228, 2011.

[90] Xochitl C Morgan, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, Samir A Shah, Neal LeLeiko, Scott B Snapper, Athos Bousvaros, Joshua Korzenik, Bruce E Sands, Ramnik J Xavier, and Curtis Huttenhower. Dysfunction of the intestinal

microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012.

[91] Akira Muto and Syozo Osawa. The guanine and cytosine content of genomic dna and bacterial evolution. volume 84, pages 166–169. National Acad Sciences, 1987.

[92] Gerard Muyzer, Ellen C De Waal, and Andre G Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16s rrna. *Applied and environmental microbiology*, 59(3):695–700, 1993.

[93] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C Linak, Aki Hirai, Hiroki Takahashi, et al. Sequence-specific error profile of illumina sequencers. *Nucleic acids research*, page gkr344, 2011.

[94] Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155–e155, 2012.

[95] H Noguchi, J Park, and T Takagi. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids research*, 2006.

[96] Mairi C Noverr, Nicole R Falkowski, Rod A McDonald, Andrew N McKenzie, and Gary B Huffnagle. Development of allergic airway disease in mice following antibiotic therapy and fungal microbiota increase: role of host genetics, antigen, and interleukin-13. *Infection and immunity*, 73(1):30–38, 2005.

[97] S M O'Mahony, G Clarke, Y E Borre, T G Dinan, and J F Cryan. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. *Behavioural brain research*, 277:32–48, January 2015.

[98] Newton G Osborne, Ralph C Wright, and Louis Grubin. Genital bacteriology: a comparative study of premenopausal women with postmenopausal women. *American journal of obstetrics and gynecology*, 135(2):195–198, 1979.

[99] Noora Ottman, Hauke Smidt, Willem M De Vos, and Clara Belzer. The function of our microbiota: who is out there and what do they do? *Frontiers in cellular and infection microbiology*, 2:104, 2012.

[100] Chana Palmer, Elisabeth M Bik, Daniel B DiGiulio, David A Relman, and Patrick O Brown. Development of the human infant intestinal microbiota. *PLoS Biol*, 5(7):e177, 2007.

[101] Kenneth D Parrish and E Peter Greenberg. A rapid method for extraction and purification of dna from dental plaque. *Applied and environmental microbiology*, 61(11):4120–4123, 1995.

[102] Helen Pearson and David Stirling. DNA Extraction from Tissue. pages 33–34. Humana Press, New Jersey, August 2003.

[103] Erica C Pehrsson, Kevin J Forsberg, Molly K Gibson, Sara Ahmadi, and Gautam Dantas. Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. *Frontiers in microbiology*, 4:145, 2013.

[104] Jacques Pépin, Nathalie Saheb, Marie-Andrée Coulombe, Marie-Eve Alary, Marie-Pier Corriveau, Simon Authier, Michel Leblanc, Geneviève Rivard, Mathieu Bettez, Valérie Primeau, Martin Nguyen, Claude-Emilie Jacob, and Luc Lanthier. Emergence of fluoroquinolones as the predominant risk factor for Clostridium difficile-associated diarrhea: a cohort study during an epidemic in Quebec. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 41(9):1254–1260, November 2005.

[105] Jane Peterson, Susan Garges, Maria Giovanni, Pamela McInnes, Lu Wang, Jeffery A Schloss, Vivien Bonazzi, Jean E McEwen, Kris A Wetterstrand, Carolyn Deal, et al. The nih human microbiome project. *Genome research*, 19(12):2317–2323, 2009.

[106] L M Proctor. The human microbiome project in 2011 and beyond. *Cell Host and Microbe*, 2011.

[107] Eamonn Martin Quigley. Gut bacteria in health and disease. *Gastroenterol hepatol (NY)*, 9(9):560–569, 2013.

[108] Andrea Raab and Jörg Feldmann. Microbial transformation of metals and metalloids. *Science progress*, 86(3):179–202, 2003.

[109] M Rho, H Tang, and Y Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic acids research*, 2010.

[110] Christian S Riesenfeld, Patrick D Schloss, and Jo Handelsman. Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics*, 38(1):525–552, December 2004.

[111] A I Rissman, B Mau, B S Biehl, A E Darling, J D Glasner, and N T Perna. Reordering contigs of draft genomes using the Mauve Aligner. *Bioinformatics*, 25(16):2071–2073, August 2009.

[112] Priyank Banthia Rita Chandki and Ruchi Banthia. Biofilms: A microbial home. *Journal of Indian Society of Periodontology*, 15(2):111, 2011.

[113] J W Rouatt and H Katznelson. A study of the bacteria on the root surface and in the rhizosphere soil of crop plants. *Journal of Applied Bacteriology*, 1961.

[114] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 1987.

[115] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[116] Elad Schiff, Neora Pick, Arie Oliven, and Majed Odeh. Multiple liver abscesses after dental treatment. *Journal of clinical gastroenterology*, 36(4):369–371, 2003.

[117] Scott Schwartz, Iddo Friedberg, Ivan V Ivanov, Laurie A Davidson, Jennifer S Goldsby, David B Dahl, Damir Herman, Mei Wang, Sharon M Donovan, and Robert S Chapkin. A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome biology*, 13(4):1, 2012.

[118] A L Servin. Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiology Reviews*, 2004.

[119] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, October 2008.

[120] Jay Shendure, Gregory J Porreca, Nikos B Reppas, Xiaoxia Lin, John P McCutcheon, Abraham M Rosenbaum, Michael D Wang, Kun Zhang, Robi D Mitra, and George M Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, September 2005.

[121] Birgitte Smith, Nan Li, Anders Schou Andersen, Hans Christian Slotved, and Karen Angeliki Krogfelt. Optimising bacterial dna extraction from faecal samples: comparison of three methods. *The open microbiology journal*, 5(1), 2011.

[122] Bärbel Stecher, Samuel Chaffron, Rina Käppeli, Siegfried Hapfelmeier, Susanne Freedrich, Thomas C Weber, Jorum Kirundi, Mrutyunjay Suar, Kathy D McCoy, Christian von Mering, Andrew J Macpherson, and Wolf-Dietrich Hardt. Like will to like: abundances of closely related species can predict susceptibility to intestinal colonization by pathogenic and commensal bacteria. *PLOS Pathogens*, 6(1):e1000711, January 2010.

[123] Bärbel Stecher and Wolf-Dietrich Hardt. The role of microbiota in infectious disease. *Trends in Microbiology*, 16(3):107–114, March 2008.

[124] E J Stewart. Growing unculturable bacteria. *Journal of Bacteriology*, 2012.

[125] Harold Swerdlow, Shaole Wu, Heather Harke, and Norman J Dovichi. Capillary gel electrophoresis for dna sequencing: laser-induced fluorescence detection with the sheath flow cuvette. *Journal of Chromatography A*, 516(1):61–67, 1990.

[126] Nobuhiro Takahashi and Bente Nyvad. The role of bacteria in the caries process ecological perspectives. *Journal of Dental Research*, 90(3):294–303, 2011.

[127] Janice E Thies. Soil microbial community analysis using terminal restriction fragment length polymorphisms. *Soil Science Society of America Journal*, 71(2):579–591, 2007.

[128] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):1, February 2012.

[129] Susannah Green Tringe, Christian Von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, et al. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, 2005.

[130] Gerardo Turcatti, Anthony Romieu, Milan Fedurco, and Ana-Paula Tairi. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research*, 36(4):e25, March 2008.

[131] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, Michael Egholm, Bernard Henrissat, Andrew C Heath, Rob Knight, and Jeffrey I Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, January 2009.

[132] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire M Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The Human Microbiome Project. *Nature*, 449(7164):804–810, October 2007.

[133] Wim H Van der Putten, John N Klironomos, and David A Wardle. Microbial ecology of biological invasions. *The ISME Journal*, 1(1):28–37, 2007.

[134] Marius Vital, Jack R Harkema, Mike Rizzo, James Tiedje, and Christina Brandenberger. Alterations of the Murine Gut Microbiome with Age and Allergic Airway Disease. *Journal of immunology research*, 2015:892568, 2015.

[135] EJ Vollaard and HA Clasener. Colonization resistance. *Antimicrobial agents and chemotherapy*, 38(3):409, 1994.

[136] Erika von Mutius and Katja Radon. Living on a farm: impact on asthma induction and clinical course. *Immunology and allergy clinics of North America*, 28(3):631–647, 2008.

[137] Alan W Walker, Jennifer Ince, Sylvia H Duncan, Lucy M Webster, Grietje Holtrop, Xiaolei Ze, David Brown, Mark D Stares, Paul Scott, Aurore Bergerat, Petra Louis, Freda McIntosh, Alexandra M Johnstone, Gerald E Lobley, Julian Parkhill, and Harry J Flint. Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *The ISME Journal*, 5(2):220–230, February 2011.

[138] Malcolm R Walter, John Bauld, and Thomas D Brock. Siliceous algal and bacterial stromatolites in hot spring and geyser effluents of yellowstone national park. *Science*, 178(4059):402–405, 1972.

[139] John B West. Regional differences in the lung. *CHEST Journal*, 74(4):426–437, 1978.

[140] William B Whitman, David C Coleman, and William J Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583, 1998.

[141] Jared Wilkening, Andreas Wilke, Narayan Desai, and Folker Meyer. Using clouds for metagenomics: a case study. pages 1–6, 2009.

[142] Benjamin P Willing, Shannon L Russell, and B Brett Finlay. Shifting the balance: antibiotic effects on host-microbiota mutualism. *Nature reviews. Microbiology*, 9(4):233–243, April 2011.

[143] Huixing Wu, Alexander Kuzmenko, Sijue Wan, Lyndsay Schaffer, Alison Weiss, James H Fisher, Kwang Sik Kim, and Francis X McCormack. Surfactant proteins a and d inhibit the growth of gram-negative bacteria by

increasing membrane permeability. *The Journal of clinical investigation*, 111(10):1589–1602, 2003.

[144] Jessica M Yano, Kristie Yu, Gregory P Donaldson, Gauri G Shastri, Phoebe Ann, Liang Ma, Cathryn R Nagler, Rustem F Ismagilov, Sarkis K Mazmanian, and Elaine Y Hsiao. Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell*, 161(2):264–276, April 2015.

[145] Pelin Yilmaz, Renzo Kottmann, Dawn Field, Rob Knight, James R Cole, Linda Amaral-Zettler, Jack A Gilbert, Ilene Karsch-Mizrachi, Anjanette Johnston, Guy Cochrane, Robert Vaughan, Christopher Hunter, Joonhong Park, Norman Morrison, Philippe Rocca-Serra, Peter Sterk, Manimozhiyan Arumugam, Mark Bailey, Laura Baumgartner, Bruce W Birren, Martin J Blaser, Vivien Bonazzi, Tim Booth, Peer Bork, Frederic D Bushman, Pier Luigi Buttigieg, Patrick S G Chain, Emily Charlson, Elizabeth K Costello, Heather Huot-Creasy, Peter Dawyndt, Todd DeSantis, Noah Fierer, Jed A Fuhrman, Rachel E Gallery, Dirk Gevers, Richard A Gibbs, Inigo San Gil, Antonio Gonzalez, Jeffrey I Gordon, Robert Guralnick, Wolfgang Hankeln, Sarah Highlander, Philip Hugenholtz, Janet Jansson, Andrew L Kau, Scott T Kelley, Jerry Kennedy, Dan Knights, Omry Koren, Justin Kuczynski, Nikos Kyrpides, Robert Larsen, Christian L Lauber, Teresa Legg, Ruth E Ley, Catherine A Lozupone, Wolfgang Ludwig, Donna Lyons, Eamonn Maguire, Barbara A Methé, Folker Meyer, Brian Muegge, Sara Nakielny, Karen E Nelson, Diana Nemergut, Josh D Neufeld, Lindsay K Newbold, Anna E Oliver, Norman R Pace, Giriprakash Palanisamy, Jörg Peplies, Joseph Petrosino, Lita Proctor, Elmar Pruesse, Christian Quast, Jeroen Raes, Sujeevan Ratnasingham, Jacques Ravel, David A Relman, Susanna Assunta-Sansone, Patrick D Schloss, Lynn Schriml, Rohini Sinha, Michelle I Smith, Erica Sodergren, Aymé Spor, Jesse Stombaugh, James M Tiedje, Doyle V Ward, George M Weinstock, Doug Wendel, Owen White, Andrew Whiteley, Andreas Wilke, Jennifer R Wortman, Tanya Yatsunenko, and Frank Oliver Glöckner. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, 29(5):415–420, May 2011.

[146] Xiaolin Tian Yung-Hua Li. Quorum Sensing and Bacterial Social Interactions in Biofilms. *Sensors (Basel, Switzerland)*, 12(3):2519, 2012.

[147] D R Zerbino and E Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 2008.

[148] Husen Zhang, John K DiBaise, Andrea Zuccolo, Dave Kudrna, Michele Braidotti, Yeisoo Yu, Prathap Parameswaran, Michael D Crowell, Rod Wing, Bruce E Rittmann, et al. Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences*, 106(7):2365–2370, 2009.

[149] Yinfeng Zhang, Cheuk-Yin Lun, and Stephen Kwok-Wing Tsui. Metagenomics: A new way to illustrate the crosstalk between infectious diseases and host microbiome. *International journal of molecular sciences*, 16(11):26263–26279, 2015.

[150] Jiangchao Zhao, Patrick D Schloss, Linda M Kalikin, Lisa A Carmody, Bridget K Foster, Joseph F Petrosino, James D Cavalcoli, Donald R VanDevanter, Susan Murray, Jun Z Li, Vincent B Young, and John J LiPuma. Decade-long bacterial community dynamics in cystic fibrosis airways. *Proceedings of the National Academy of Sciences of the United States of America*, 109(15):5809–5814, April 2012.

[151] W Zhu, A Lomsadze, and M Borodovsky. Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 2010.

Appendix

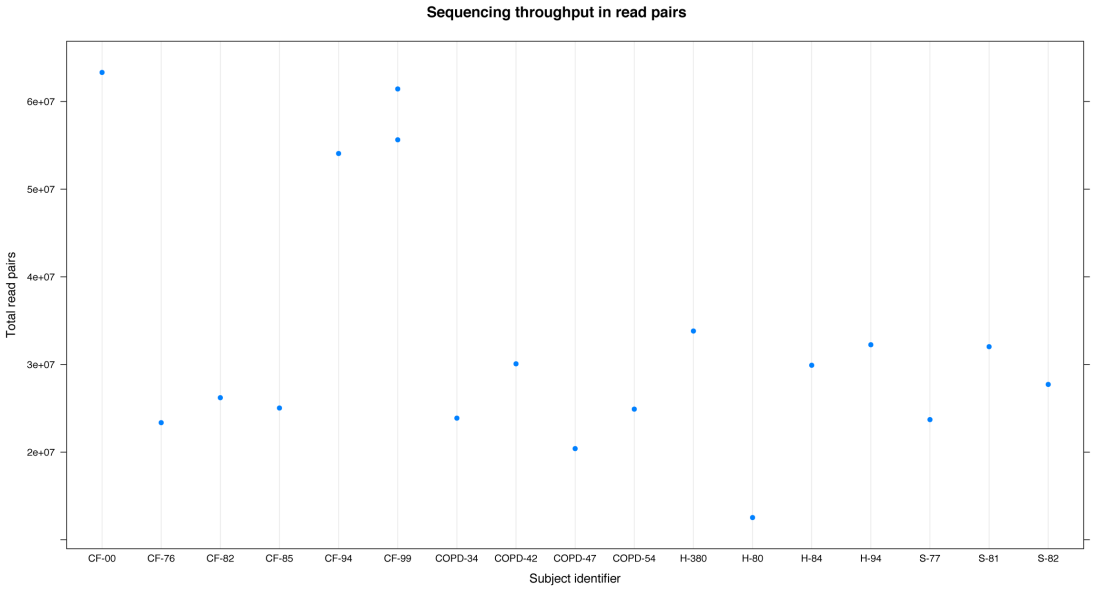

Figure 7.1: Plot showing the sequencing throughput for each sample. Subject ID on x-axis and total number of read pairs on y-axis. CF-99 was sampled twice.

Patient ID	Clinically detected bacteria	confirmed using genomic analysis
CF-99_2	Staphylococcus aureus	yes
CF-99	Staphylococcus aureus	yes
CF-76	Pseudomonas aeruginosa	yes
CF-85	Achromobacter Xylosoxidans	yes
	Staphylococcus aureus	yes
CF-82	Pseudomonas aeruginosa	yes
CF-94	Staphylococcus aureus	no
CF-00	Stenotrophomonas maltophilia	yes
	Haemophilus influenzae	yes

Table 7.1: Table showing clinically detected bacteria and their subsequent identification using genomic analysis

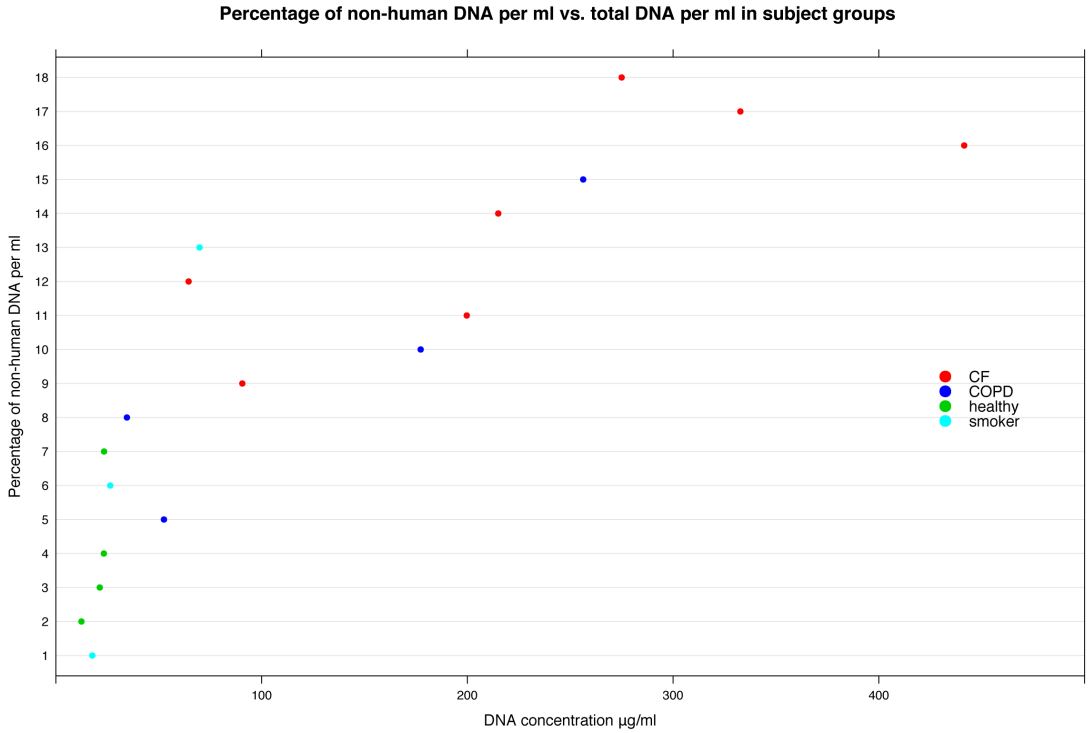

Figure 7.2: Plot showing the percentage of non-human DNA per ml vs. total DNA concentration. CF samples are marked in red, COPD in blue, healthy in green and smokers in light blue.

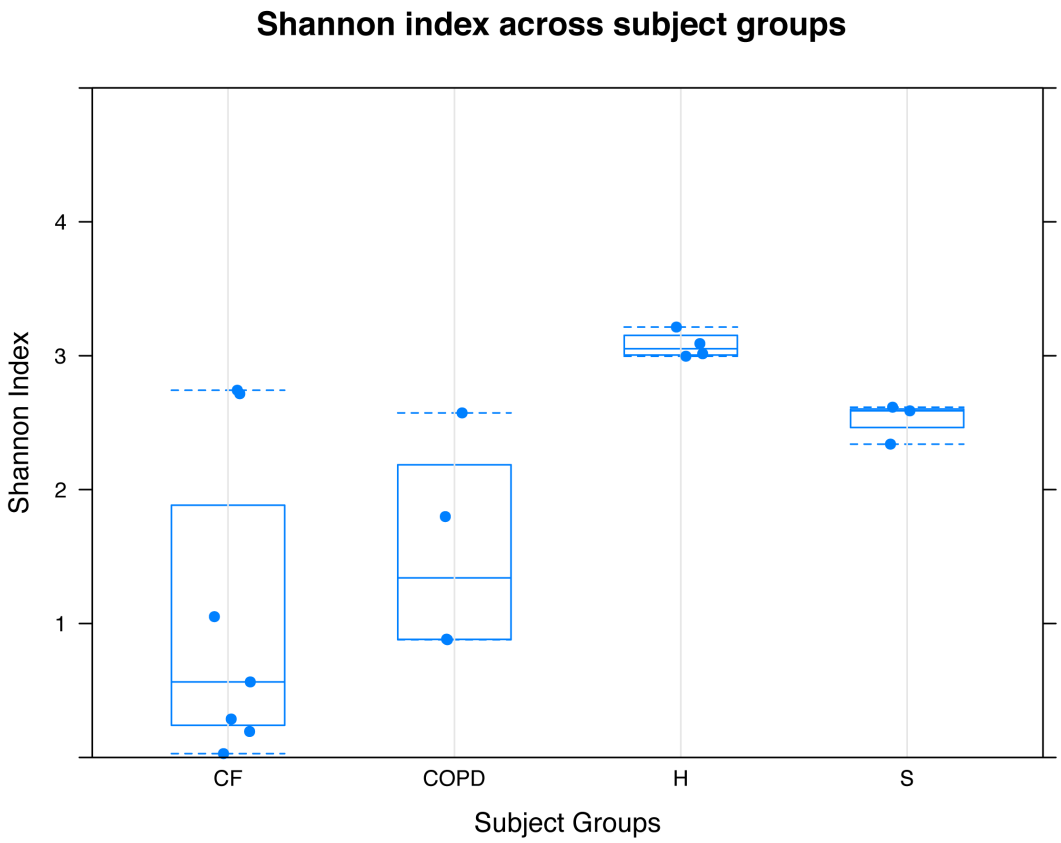

Figure 7.3: Plot showing the Shannon entropy of each sample from the four subject groups.

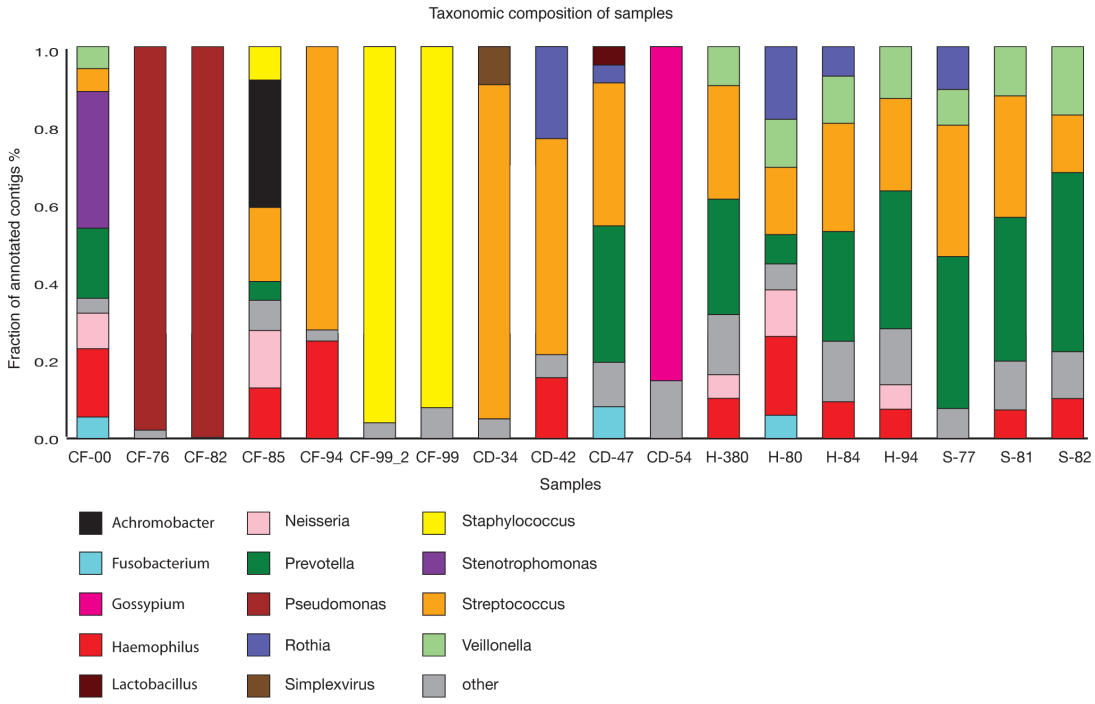

Figure 7.4: Genus level taxonomic composition of each sample. All genera constituting less than 4.5% of the annotated fraction have been labeled as others.

	CF-85	CF-82	CD-42	CF-76	CF-00	CD-47	CD-34	CD-54	CF-94	CF-99_2	CF-99
Date of Sputum	03.07.2012	10.07.2012	20.07.2012	04.07.2012	14.03.2013	10.07.2012	08.08.2012	08.08.2012	08.08.2012	23.02.2013	15.05.2013
QUANT(mg)	780	3560	1220	180	1100	420	700	1400	760	370	540
YOB	1985	1982	1942	1976	2000	1947	1934	1954	1994	1999	1999
SEX	f	m	f	f	m	m	m	f	f	f	f
GROUP	CF	CF	COPD	CF	CF	COPD	COPD	COPD	CF	CF	CF
Antibiotics	Dalacin	Ciprofloxacin	None	Tazobac + Ciprofloxacin	Amoxicillin/Clavulanic-Acid	None	Augmentin	Augmentin	Amoxicillin/ Clavulanic-Acid	TMP/SMX	Clarithromycin
Date of Antibiotics*	03.07.2012	13.04.2011	NA	11.08.2010	14.3.2013	NA	08.08.2012	06.08.2012	8.8.2012	23.2.2013	16.4.2013
FEV1	1.09 L (37% Soll)	1.36 L (33% Soll)	0.7 L (32% Sc	1.71 L (63% Soll)	1.65L, 98%	1.87 L (57% Soll)	0.71 L (23% Soll)	0.58 L (23% Soll)	1.6L, 54%	2.9L, 85%	3.07L, 88%
Date of FEV1	18.06.2012	25.01.2012	17.07.2012	13.03.2012	14.3.2013	18.05.2012	13.08.2012	03.05.2012	2.8.2012	23.2.2013	15.5.2013
Date of Exacerbation	04.06.2012	23.11.2010	NA	11.08.2010	uk/01/2013	NA	07.08.2012	06.08.2012	uk/07/2012	29.1.2013	NA

Table 7.2: Table showing the CF and COPD patient demographics data

CF-76		
Clinically Tested Antibiotcs	Clinical Results: 26.06.2012	Genomic Results: 04:07.2012
Pip-Tazobactam	Sensitive	No resistance detected
Caftazidim	Sensitive	No resistance detected
cefepime	Sensitive	No resistance detected
Aztreonam	Sensitive	No resistance detected
Imipenem	Sensitive	No resistance detected
Meropenem	Sensitive	No resistance detected
Amikacin	Sensitive	No resistance detected
Netilmicin	Sensitive	No resistance detected
Tobramycin	Sensitive	No resistance detected
Ciprofloxacin	Sensitive	No resistance detected
Colistin	Sensitive	No resistance detected

Table 7.3: clinically tested antibiotics for detecting resistance and results from genomics analysis for CF-76

CF-00		
Clinically Tested Antibiotics	Clinical Results: 14.03.2013	Genomic Results: 14.03.2013
Ampicillin	Sensitive	No resistance detected
Amoxicillin-Calv.	Sensitive	No resistance detected
Cefuroxim	Sensitive	No resistance detected
Ceftriaxon	Sensitive	No resistance detected
Meropenem	Sensitive	No resistance detected
Ciprofloxacin	Sensitive	No resistance detected
Co-Trimoxazol	Sensitive	No resistance detected

Table 7.4: clinically tested antibiotics for detecting resistance and results from genomics analysis for CF-00

CF-82		
Clinically Tested Antibiotcs	Clinical Results: 29.04.2012	Genomic Results: 10.07.2012
Pip Tazobactam	Sensitive	No resistance detected
Ceftazidim	Sensitive	No resistance detected
Cefepime	Sensitive	No resistance detected
Aztreonam	Sensitive	No resistance detected
Imipenem	Sensitive	No resistance detected
Meropenem	Sensitive	No resistance detected
Amikacin	Sensitive	No resistance detected
Netilmicin	Sensitive	No resistance detected
Tobramycin	Sensitive	No resistance detected
Ciprofloxacin	Intermediate	Resistant
Colistin	Sensitive	Resistant

Table 7.5: clinically tested antibiotics for detecting resistance and results from genomics analysis for CF-82

CF-94		
Clinically Tested Antibiotics	Clinical Results: 06.08.2012	Genomic results:08.08.2012
Penicillin G	Resistant	No resistance detected
Ampicillin	Resistant	No resistance detected
Oxacillin	Sensitive	No resistance detected
Amoxicillin-Clav.	Sensitive	No resistance detected
Cefazolin	Sensitive	No resistance detected
Cefamandol	Sensitive	No resistance detected
Cefuroxim	Sensitive	No resistance detected
Imipenem	Sensitive	No resistance detected
Meropenem	Sensitive	No resistance detected
Rifampicin	Sensitive	No resistance detected
Ciprofloxacin	Sensitive	No resistance detected
Co-Trimoxazol	Sensitive	No resistance detected
Clindamycin	Sensitive	No resistance detected
Azithromycin	Sensitive	No resistance detected
Clarithromycin	Sensitive	No resistance detected
Erythromycin	Sensitive	No resistance detected
Tetracyclin	Sensitive	No resistance detected
Fusidinsäure	Sensitive	No resistance detected

Table 7.6: clinically tested antibiotics for detecting resistance and results from genomics analysis for CF-94

CF-99		
Clincally Tested Antibiotics	Clinical Results: 15.05.2013	Genomic Results: 15.05.2013
Penicillin G	Resistant	No resistance detected
Ampicillin	Resistant	No resistance detected
Oxacillin	Sensitive	No resistance detected
Amoxicillin-Clav.	Sensitive	No resistance detected
Pip-Tazobactam	Sensitive	No resistance detected
Cefamandol	Sensitive	No resistance detected
Cefuroxim	Sensitive	No resistance detected
Imipenem	Sensitive	No resistance detected
Meropenem	Sensitive	No resistance detected
Gentamicin	Sensitive	No resistance detected
Rifampicin	Sensitive	No resistance detected
Ciprofloxacin	Resistant	No resistance detected
levofloxacin	Sensitive	No resistance detected
Co-Trimoxazol	Sensitive	No resistance detected
Clindamycin	Sensitive	No resistance detected
Azithromycin	Sensitive	No resistance detected
Clarithromycin	Sensitive	No resistance detected
Erythromycin	Sensitive	No resistance detected
Tetracyclin	Sensitive	No resistance detected
Fusidinsäure	Sensitive	No resistance detected

Table 7.7: clinically tested antibiotics for detecting resistance and results from genomics analysis for CF-99

CF-99_2		
Clinically Tested Antibiotics	Clinical Results:23.02.2012	Genomic results:23.02.2012
Penicillin G	Resistant	No resistance detected
Ampicillin	Resistant	No resistance detected
Oxacillin	Sensitive	No resistance detected
Amoxicillin-Clav.	Sensitive	No resistance detected
Cefazolin	Sensitive	No resistance detected
Cefamandol	Sensitive	No resistance detected
Cefuroxim	Sensitive	No resistance detected
Imipenem	Sensitive	No resistance detected
Meropenem	Sensitive	No resistance detected
Gentamicin	Sensitive	No resistance detected
Tobramycin	Sensitive	No resistance detected
Rifampicin	Sensitive	No resistance detected
Ciprofloxacin	Sensitive	No resistance detected
Co-Trimoxazol	Sensitive	No resistance detected
Clindamycin	Sensitive	No resistance detected
Azithromycin	Sensitive	No resistance detected
Clarithromycin	Sensitive	No resistance detected
Erythromycin	Sensitive	No resistance detected
Tetracyclin	Sensitive	No resistance detected
Fusidinsäure	Sensitive	No resistance detected

Table 7.8: clinically tested antibiotics for detecting resistance and results from genomics analysis for CF-992

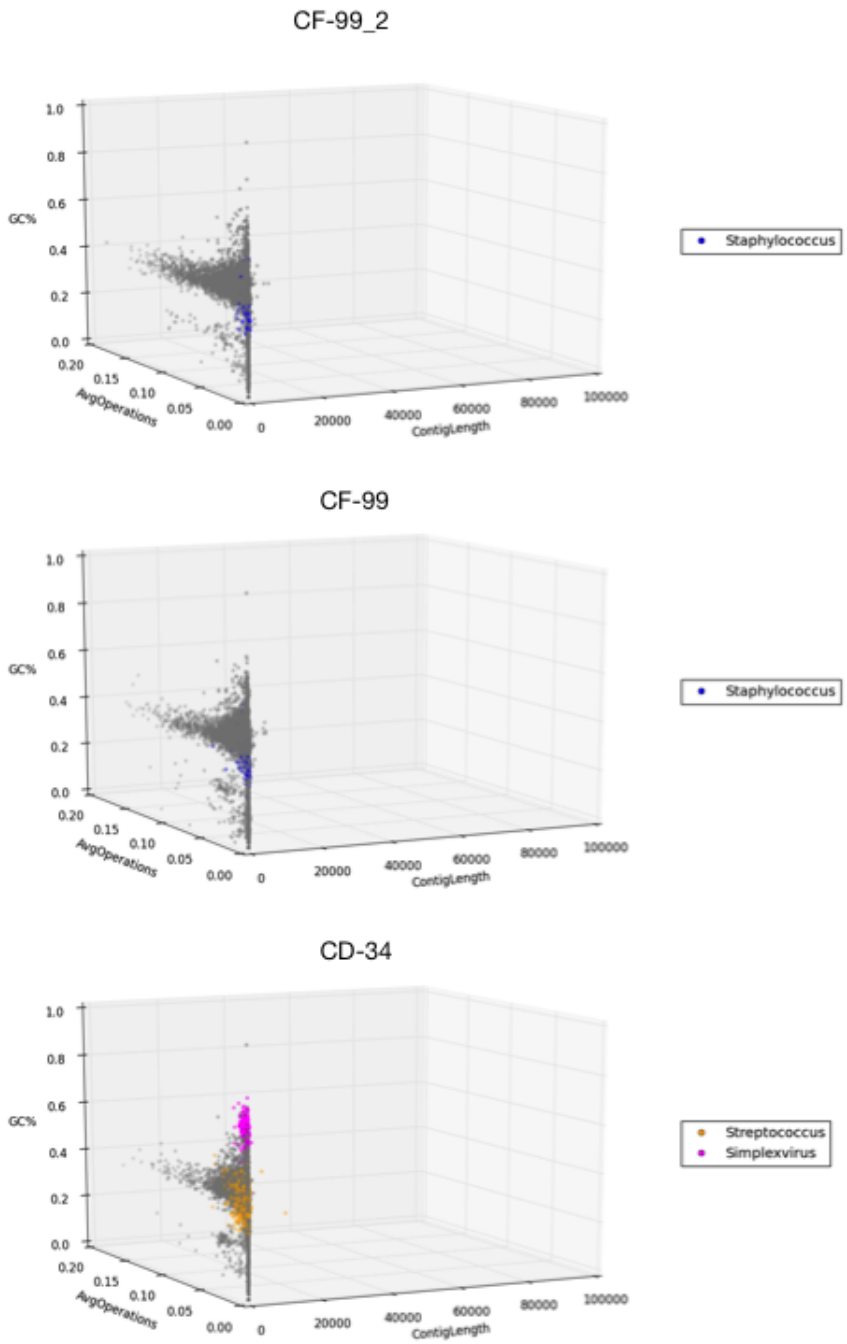

Figure 7.5: Entropy Landscape plot showing lung microbial composition for CF-99 and CD-34

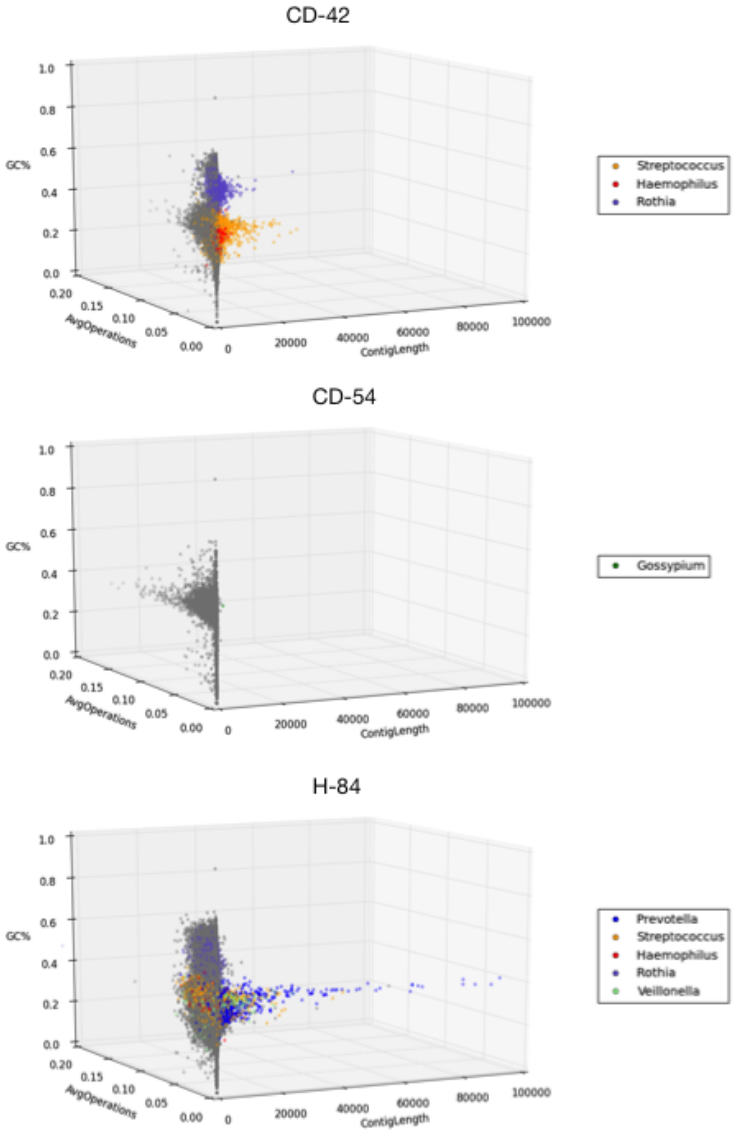

Figure 7.6: Entropy Landscape plot showing lung microbial composition for CD-42, CD-54, H-84

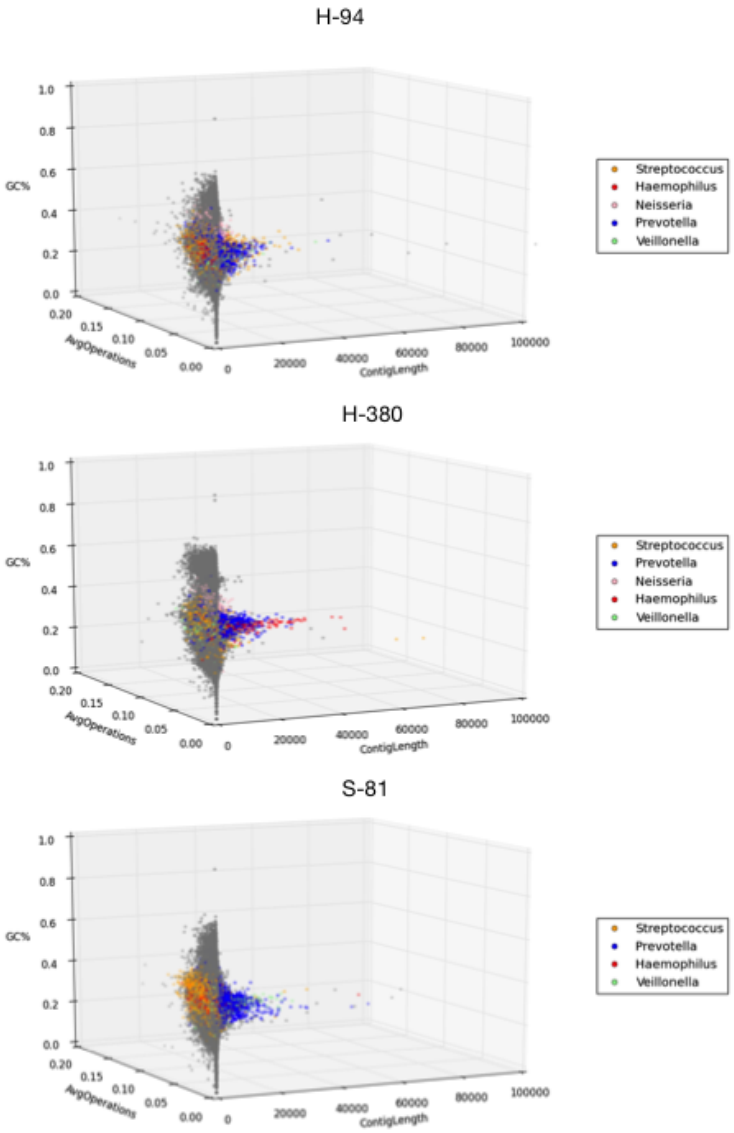

Figure 7.7: Entropy Landscape plot showing lung microbial composition for H-380, H-94, S-81

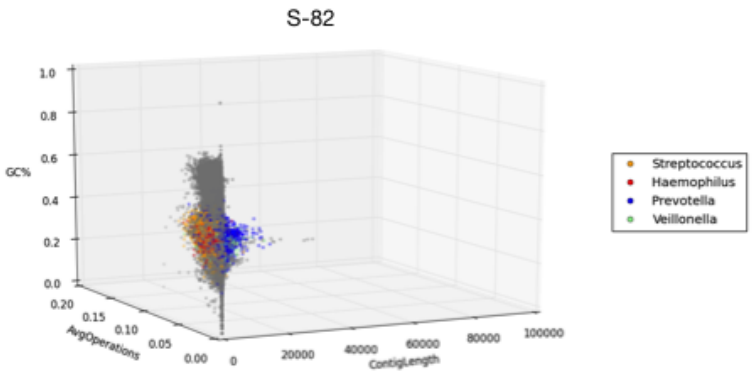

Figure 7.8: Entropy Landscape plot showing lung microbial composition for S-82

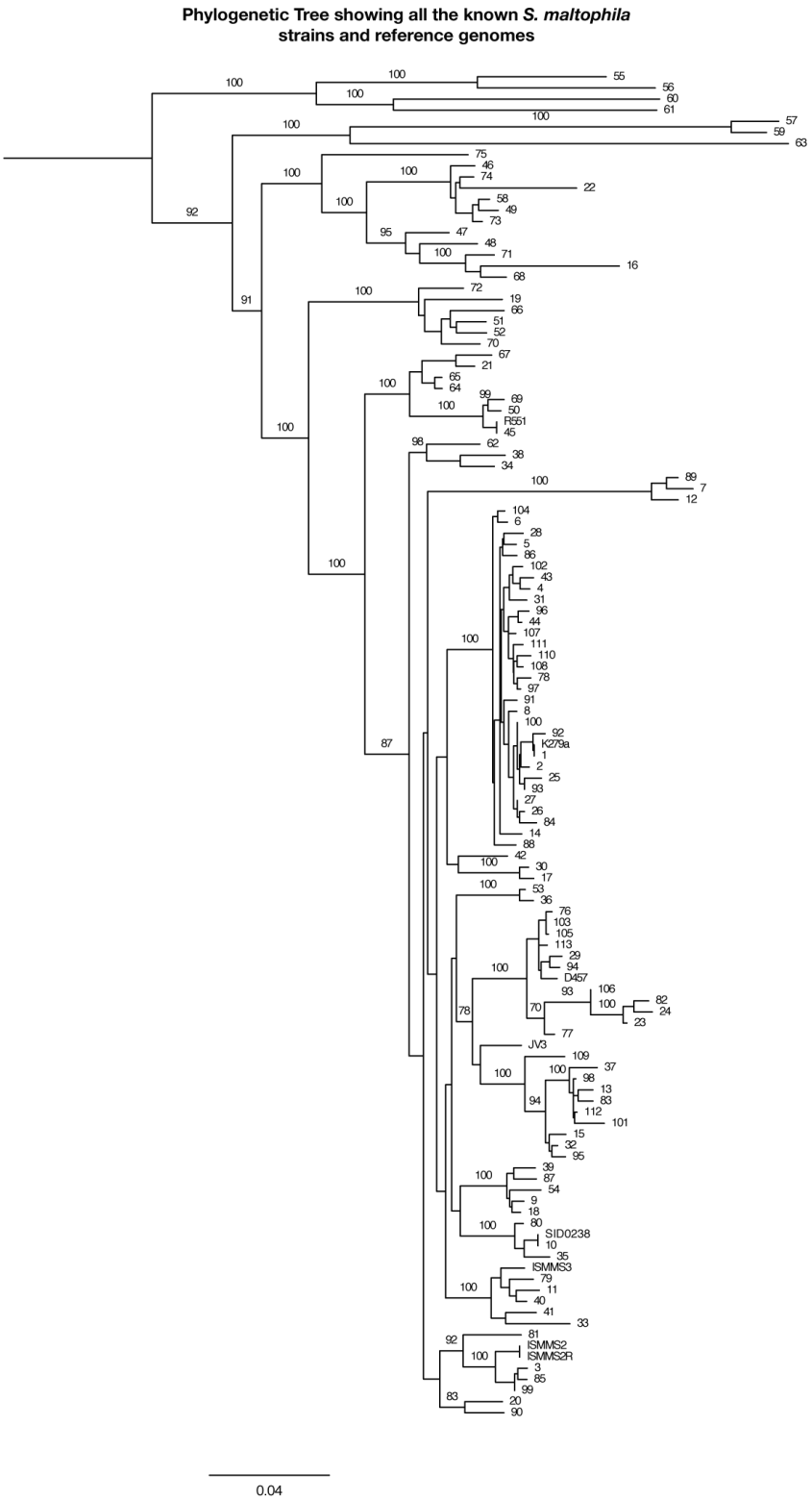

Figure 7.9: Phylogenetic tree showing placement of reference *S. maltophilia* database strains in addition to the strain isolated from our subject CF-00 and all strains documented in the PUBMLST database.

Acknowledgements

Firstly, I would like to thank Prof. Christian von Mering, my supervisor for this excellent opportunity to work in the new and budding field of human microbial metagenomics. Christian has always given me extensive freedom to design my own projects and has extended his immense support in carrying out my research. His enthusiastic and innovative attitude has been a source of encouragement whenever I faced a setback. He has been an inspiring mentor during the past four years.

I am grateful to my thesis committee members Prof. Shimizu, Prof. Hardt and former member Prof. Held for their time, constructive criticism and suggestions on my work. I would also like to thank my colleagues from the Mering lab for the motivating environment, especially Dr. Joao Rodrigues for helpful brain storming sessions. Thank you to the administrative staff, especially Dr. Werner Wolz, Dr. Susanna Bachmann and Mrs. Ursula Witschi for their prompt support whenever needed. I would also like to thank the functional genomics facility for providing easy access to world class sequencing technologies and facilitating my reseach.

Thank you to all my research collaborators for letting me be a part of such a diverse array of projects. I have truly enjoyed and learned a lot form being a part of many talented and passionate scientific teams. The knowledge transfer has remarkably benefitted me.

Lastly, I would like to thank my family, especially my husband for his unwavering support. I would like to dedicate this work to my husband, Justin Feigelman who has provided me with the strength and guidance that I needed to complete my PhD endeavors.